



The Open
University

M249 Practical modern
statistics

Computer Book 2

Time series

About this module

M249 *Practical modern statistics* uses the software packages *IBM SPSS Statistics* (SPSS Inc.) and *WinBUGS*, and other software. This software is provided as part of the module, and its use is covered in the *Introduction to statistical modelling* and in the four computer books associated with *Books 1 to 4*. This computer book contains all the computer work associated with *Book 2*.

Cover image courtesy of NASA. This photograph, acquired by the ASTER instrument on NASA's Terra satellite, shows an aerial view of a large alluvial fan between the Kunlun and Altun mountains in China's Xinjiang province. For more information, see NASA's Earth Observatory website at <http://earthobservatory.nasa.gov>.

This publication forms part of an Open University module. Details of this and other Open University modules can be obtained from the Student Registration and Enquiry Service, The Open University, PO Box 197, Milton Keynes MK7 6BJ, United Kingdom (tel. +44 (0)845 300 60 90; email general-enquiries@open.ac.uk).

Alternatively, you may visit the Open University website at www.open.ac.uk where you can learn more about the wide range of modules and packs offered at all levels by The Open University.

To purchase a selection of Open University materials visit www.ouw.co.uk, or contact Open University Worldwide, Walton Hall, Milton Keynes MK7 6AA, United Kingdom for a brochure (tel. +44 (0)1908 858779; fax +44 (0)1908 858787; email ouw-customer-services@open.ac.uk).

The Open University, Walton Hall, Milton Keynes MK7 6AA.

First published 2007. Second edition 2013.

Copyright © 2007, 2013 The Open University

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, transmitted or utilised in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without written permission from the publisher or a licence from the Copyright Licensing Agency Ltd. Details of such licences (for reprographic reproduction) may be obtained from the Copyright Licensing Agency Ltd, Saffron House, 6–10 Kirby Street, London EC1N 8TS (website www.cla.co.uk).

Open University materials may also be made available in electronic formats for use by students of the University. All rights, including copyright and related rights and database rights, in electronic materials and their contents are owned by or licensed to The Open University, or otherwise used by The Open University as permitted by applicable law.

In using electronic materials and their contents you agree that your use will be solely for the purposes of following an Open University course of study or otherwise as licensed by The Open University or its assigns.

Except as permitted above you undertake not to copy, store in any medium (including electronic storage or use in a website), distribute, transmit or retransmit, broadcast, modify or show in public such electronic materials in whole or in part without the prior written consent of The Open University or in accordance with the Copyright, Designs and Patents Act 1988.

Edited, designed and typeset by The Open University, using the Open University T_EX System.

Printed in the United Kingdom by The Charlesworth Group, Wakefield.

ISBN 978 1 7800 7662 1

Contents

Chapter 1	Exploring time series	4
1.1	Entering time series data	4
1.2	Displaying time series	6
1.3	Transforming time series	8
Chapter 2	Moving averages	12
Chapter 3	Estimating the components of a time series	16
Chapter 4	Simple exponential smoothing	21
Chapter 5	Holt’s and Holt–Winters exponential smoothing	26
Chapter 6	Correlograms and forecasting	33
Chapter 7	Stationary time series	38
Chapter 8	ARIMA modelling in SPSS	42
8.1	Selecting an ARIMA model	42
8.2	Fitting and checking ARIMA models	45
8.3	Forecasting with ARIMA models	53
Computer Exercises on Book 2		56
Learning outcomes		58
Solutions to Computer Activities		59
Solutions to Computer Exercises		79
Index		86

Chapter 1

Exploring time series

The additive decomposition model for a time series was described in Section 2 of *Book 2*. You learned how to decide whether or not an additive model might be appropriate for a time series by examining the time plot for the series. In this chapter, you will learn how to enter time series data in SPSS (in Section 1.1), and how to produce time plots for a time series (Section 1.2) and for a log transformed series (Section 1.3). In Section 1.3, you will also learn how to transform a variable using a power transformation, so that you can investigate whether an additive model might be appropriate for the transformed series.

1.1 Entering time series data

Much of the process of entering time series data in SPSS is the same as for entering other types of data. However, SPSS has a special way of defining dates for time series. This is described in Computer Activity 1.1.

Computer Activity 1.1 Entering time series data and defining dates

The annual average temperatures (in degrees Celsius) in Central England for each year from 1970 to 2004 are given in Table 1.1.

Table 1.1 Annual average temperatures in Central England, 1970–2004

9.57	9.68	9.19	9.54	9.62	10.00	10.08	9.48	9.42	8.81
9.42	9.24	9.84	10.03	9.73	8.86	8.74	9.05	9.77	10.50
10.63	9.52	9.86	9.49	10.24	10.52	9.20	10.53	10.34	10.63
10.30	9.93	10.60	10.50	10.49					

The data in the first row of the table are for the years 1970 to 1979, those in the second row are for 1980 to 1989, and so on. Note that the years are not given explicitly in the table. In this activity, you will learn how to enter the years in SPSS using a special facility for defining dates.

Run SPSS now. Make sure that the **Data View** panel is uppermost in the **Data Editor**.

Enter the data from Table 1.1 in the first column of the data sheet in the following order: 9.57, 9.68, 9.19, ... When you have entered the 35 values, check them against Table 1.1.

Notice that the variable name in the column heading has changed to **VAR00001**. Change the variable name to **temperature**, as follows.

- Click on the **Variable View** tab near the bottom left-hand corner of the window so that the **Variable View** panel is uppermost.
- Replace the default variable name **VAR00001** by typing **temperature** in its place.
- If necessary, widen the **Name** column so that you can read the full variable name.

Now that you have entered the temperatures for 1970 to 2004, the next stage is to enter the years. You could do this by simply typing them in. However, when time points are equally spaced, as is the case for time series, the times (or dates) can be entered more conveniently using **Define Dates...** from the **Data** menu. All you need to do is specify the units (years, months, etc.) and the first value.

This data set is a subset of the data set introduced in Example 1.2 of *Book 2*.

If the **Data View** panel is not uppermost, then click on the **Data View** tab in the lower left-hand corner of the window.

Place the mouse pointer on the vertical line separating the column headings **Name** and **Type**, hold the mouse button down, and drag the mouse to adjust the width of the **Name** column.

Enter the years, as follows.

- ◇ Choose **Define Dates...** from the **Data** menu. The **Define Dates** dialogue box will open.
- ◇ Scroll to the top of the list of options in the **Cases Are** area on the left-hand side of the dialogue box.

The first option in the list is **Years**. This option is appropriate when the data represent years, as is the case here. The second entry is **Years, quarters**. This is appropriate when the data relate to quarters within years — for example, quarter 3, 2005, quarter 4, 2005, quarter 1, 2006, and so on. Scroll through the list to see what options are available.

- ◇ Select **Years** (by clicking on it).
- ◇ Enter 1970 in the **Year** field that appears in the **First Case Is** area.
- ◇ Click on **OK**.

The **Viewer** window will open and inform you that two new variables have been created: **YEAR_** and **DATE_**. Look at these variables in the **Data Editor**. In **Data View**, you will see the two new variables, both of which list the years 1970 to 2004. In **Variable View**, notice that **YEAR_** is a numeric variable, whereas **DATE_** is a string variable. (A numeric variable contains numbers which can be used in calculations. Values of a string variable contain characters, which may include numbers, but they cannot be used in arithmetic calculations.)

The variable names **YEAR_** and **DATE_**, complete with underscores, are recognized as special date variables by SPSS, so they should not be changed. Their labels, which are in the **Label** column, are also used by some SPSS routines, so you should not change them either.

This completes the data entry. You may save your work if you wish. However, you do not need to as the data have already been saved in a file named **temperature1.sav**. This file is located in the **Book 2** subfolder of the **M249 Data Files** folder in **Documents**.

All the data files used in this computer book are located in the **Book 2** subfolder of the **M249 Data Files** folder.

Computer Activity 1.2 Monthly temperature data

In this activity, you will use **Define Dates...** to define the dates for a monthly time series. Open the SPSS data file **temperature2.sav**. There is one variable, **temperature**, which lists the monthly average temperatures in Central England between January 1970 and December 2004. Create the date variables, as follows.

- ◇ Obtain the **Define Dates** dialogue box.
- ◇ Select **Years, months** in the list of options in the **Cases Are** area.
- ◇ In the **First Case Is** area, enter 1970 in the **Year** field, and 1 in the **Month** field (1 stands for January, 2 for February, and so on).
- ◇ Click on **OK**.

Data > Define Dates...

Three new variables will be created, named **YEAR_**, **MONTH_** and **DATE_**. Look at these variables in the **Data View** panel of the **Data Editor**. **YEAR_** gives the year (1970 to 2004), **MONTH_** gives the month (1 to 12) and **DATE_** combines them in the form JAN 1970, FEB 1970, and so on.

Now look at the variable labels in the **Variable View** panel. SPSS tends to give detailed labels to the date variables it creates. The variable **MONTH_** is labelled **MONTH, period 12**. This means that SPSS has recognized that the months repeat in a cycle of period 12.

These data have been saved in the file **temperature3.sav**.

Using **Define Dates...** saves you work as you do not have to type in all the values of the time variable. Also, if appropriate, the periodicity of the time variable is defined automatically for use with other SPSS commands. In this computer book, all seasonality is annual, and only monthly and quarterly time series are used with **Define Dates...** Computer Activity 1.3 will give you some practice in using **Define Dates...** for a quarterly time series.

Computer Activity 1.3 Beer consumption

The SPSS data file **beerquarter1.sav** contains the time series of quarterly beer consumption in the UK, from the first quarter of 1991 to the second quarter of 2004. Open the file now. There is one variable, **consumption**; this contains the beer consumption in thousands of hectolitres for each quarter.

These data are described in Activity 1.3 of *Book 2*.

Use **Define Dates...** to create the time variables for this time series: use the **Years, quarters** option in the **Cases Are** area of the **Define Dates** dialogue box. Examine the time variables in the **Data Editor**. Check that the newly defined time variables begin with the first quarter of 1991 and end with the second quarter of 2004, and that SPSS has correctly recognized that the period is 4.

These data with defined dates have been saved as **beerquarter2.sav**.

1.2 Displaying time series

Time plots are an important tool for describing time series and making an initial model choice. They are produced using **Sequence Charts...** from the **Forecasting** submenu of **Analyze**. The method is described in Computer Activity 1.4.

Computer Activity 1.4 Obtaining a time plot

The SPSS data file **infections.sav** contains the time series of weekly reports from the UK of four infectious agents. The series begins in mid-1996 and ends in mid-2003. Open the file now.

Some of these data are discussed in Example 2.2 of *Book 2*.

There are seven variables. The first three of these are time variables: **year**, **week** (numbered 1 to 52) and **date** (in week and year format). These variables were not created using **Define Dates...**: since a year contains one or two days more than 52 weeks, the **Define Dates** dialogue box does not have a **Year, week** option.

The next four variables are counts of reports of four infectious agents: **clostridium** (*Clostridium difficile*), **rsv** (Respiratory Syncytial virus, which causes a 'flu-like' illness), **rotavirus**, and **salmonella** (*Salmonella Typhimurium* DT104).

The reports for the one or two extra days in each year were deleted when constructing this time series.

Obtain a time plot of the time series of Salmonella reports, as follows.

- ◇ Choose **Sequence Charts...** from the **Forecasting** submenu of **Analyze**. The **Sequence Charts** dialogue box will open.
- ◇ Enter **salmonella** in the **Variables** field.
- ◇ Enter **date** in the **Time Axis Labels** field. (Note that the variable **date** is referred to by its label, **Date**.)
- ◇ Click on **OK**.

The time plot shown in Figure 1.1 will be displayed in the **Viewer** window.

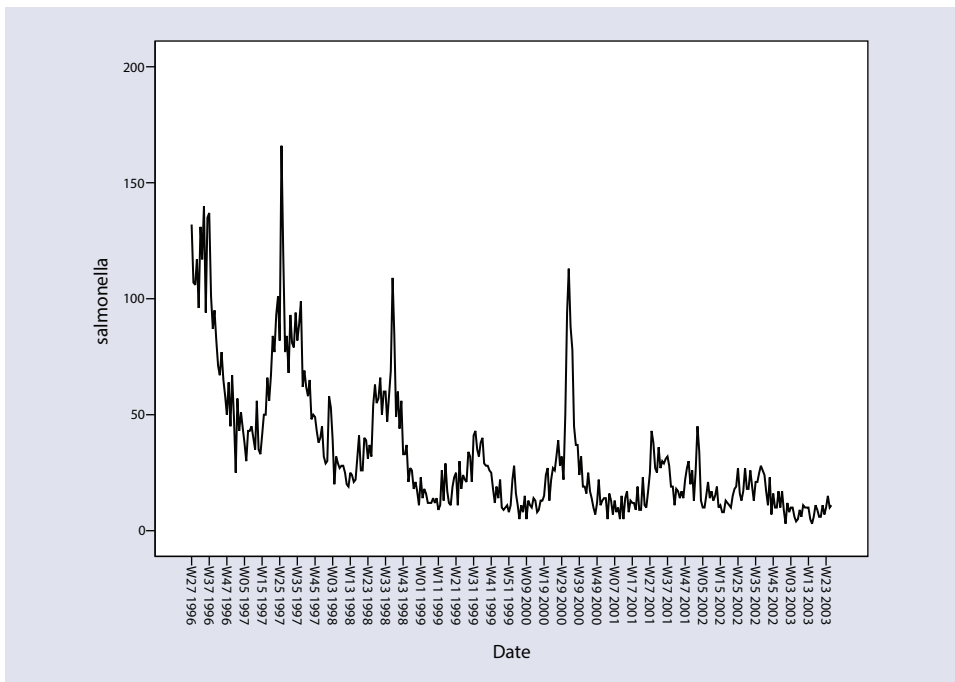


Figure 1.1 Default time plot of Salmonella reports

In Figure 1.1, there is a label on the horizontal axis every 10 weeks. This makes it hard to identify seasonal patterns. The labelling of the time axis can be changed using the **Chart Editor**. Edit the labelling so that there is a label every 26 weeks, as follows.

- ◇ Place the mouse pointer on the graph, and double-click to open the **Chart Editor**.
- ◇ Click on the large **X** in the main toolbar to open the **Properties** dialogue box.
- ◇ If necessary, click on the **Labels & Ticks** tab to bring the **Labels & Ticks** panel uppermost.
- ◇ In the **Major Increment Labels** area, choose **Diagonal** from the **Label orientation** drop-down list. (This changes the orientation of the labels.)
- ◇ Under **Category Label Placement**, select **Custom**.
- ◇ Enter 25 in the **Ticks skipped between labels** field. With this setting, there will be a tick and label on the time axis for every 26th time point.
- ◇ Click on **Apply**, then click on **Close** to close the **Properties** dialogue box.
- ◇ Close the **Chart Editor**.

The edited time plot in the **Viewer** window will be as shown in Figure 1.2.

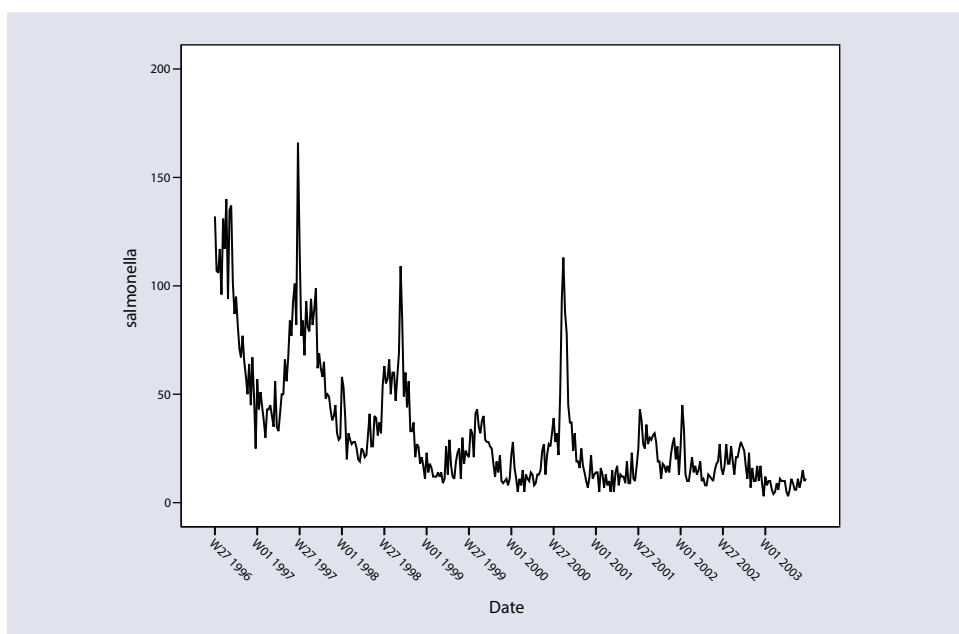


Figure 1.2 Time plot of Salmonella reports, with edited labelling

Notice that time labels are given for week 1 and week 27 of each year. This spacing of the labels makes it easier to see whether or not the time series is seasonal. The peaks occur close to week 27 each year, so the time series is indeed seasonal, with a summer peak.

It is often useful to experiment with the orientation and spacing of the labels in this way in order to obtain a time plot that is as informative as possible. Editing the labelling can also improve the appearance and readability of a time plot. This has been done in many of the time plots reproduced in this computer book. You are encouraged to do likewise when producing time plots, even when not explicitly asked to: it is good practice.

Note that SPSS may not allow you complete flexibility in choosing the labels to be displayed if your choices would produce labels that are too crowded.

Do not close the file **infections.sav**. You will need it for the next computer activity.

1.3 Transforming time series

When a time series cannot be described adequately by an additive model, it is sometimes possible to find a transformation such that an additive model is appropriate for the transformed time series. In particular, as you saw in Subsection 2.3 of *Book 2*, if a multiplicative model is appropriate for a time series, then transforming the data by taking logarithms will produce a series for which an additive model is suitable. The time plot of a log transformed series can be obtained using **Sequence Charts...** from the **Forecasting** submenu of **Analyze**. This is illustrated in Computer Activity 1.5.

Computer Activity 1.5 Log transformations

- In *Book 2*, it was suggested that the time series of weekly rotavirus infections might be described better by a multiplicative model than by an additive model. Data for 1996 to 1998 were used there. Data for 1996 to 2003 are in the SPSS data file **infections.sav**. In this activity, you will use these data to investigate whether a multiplicative model is also better for this longer time series. (Open this file now, if it is not already open.)
- Obtain the **Sequence Charts** dialogue box.
 - Enter **rotavirus** in the **Variables** field, and **date** in the **Time Axis Labels** field.
 - Click on **OK**.

The time plot that will be displayed in the **Viewer** window is shown in Figure 1.3(a).

See Example 2.3 of *Book 2*.

Analyze > Forecasting > Sequence Charts...
You may need to remove salmonella from the Variables field.

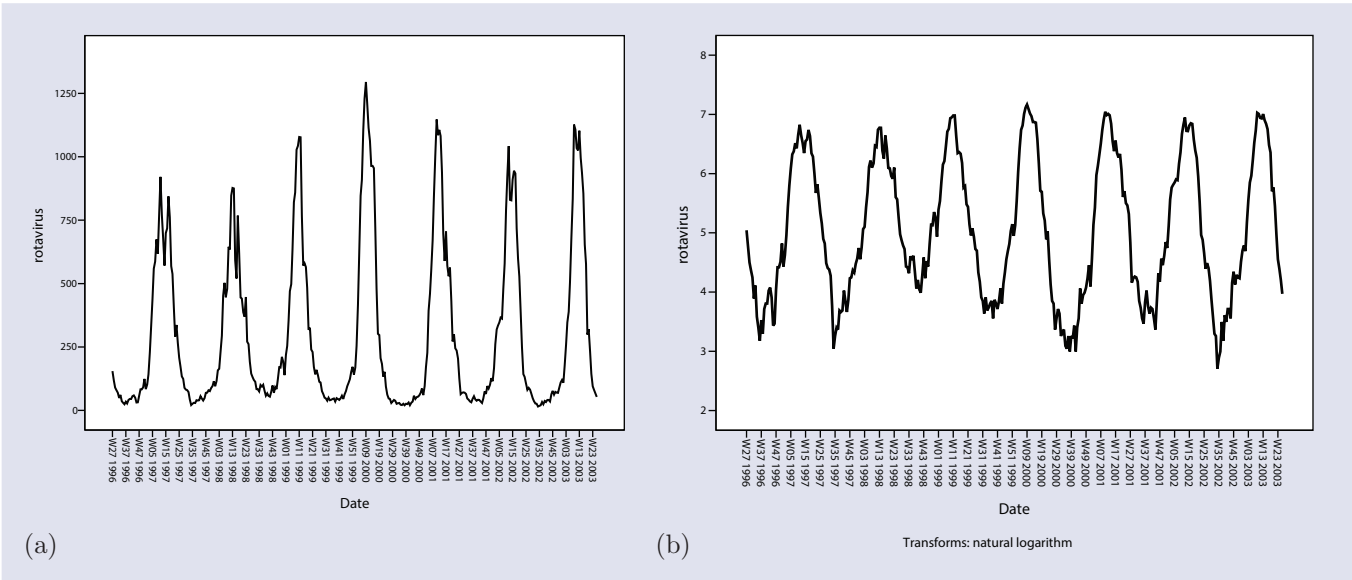


Figure 1.3 Time plots of rotavirus infections: (a) counts (b) logarithms of counts

- Now obtain the time plot of the logarithms of the weekly rotavirus infection counts, as follows.
- Obtain the **Sequence Charts** dialogue box. The settings you used to obtain the time plot of counts will have been retained.
 - In the **Transform** area, check the **Natural log transform** check box. Leave all the other boxes unchecked.
 - Click on **OK**.

Analyze > Forecasting > Sequence Charts...

The time plot that will be displayed in the **Viewer** window is shown in Figure 1.3(b). The size of the irregular fluctuations in this time plot appears to be similar at the peaks and in the troughs of the seasonal cycle. This suggests that an additive model may be appropriate for the log transformed series. Equivalently, a multiplicative model may be appropriate for the original time series.

Note that when using the **Sequence Charts** dialogue box, the log transformed values are not displayed in the data sheet. If required, these values can be obtained using **Compute Variable...** from the **Transform** menu, as described in the *Introduction to statistical modelling*.

Computer Activity 1.6 Airline passengers

The SPSS data file **airline.sav** contains the time series of monthly total numbers of UK airline passengers, from January 1974 to December 1999. Open the file now. There are four variables: **passengers**, which contains the total number of passengers, **YEAR_**, **MONTH_** and **DATE_**.

- (a) Obtain the time plot of **passengers** by **DATE_**. (Note that within the **Sequence Charts** dialogue box, **DATE_** is referred to by its label, **Date**.) Describe the main features of the time plot.
- (b) Edit the labelling of the time axis so that labels are displayed diagonally and are included only for January of each year. What do you conclude about seasonal variation?
- (c) Obtain a time plot of the logarithms of numbers of passengers.
- (d) Is an additive model appropriate for the original (untransformed) time series? Is a multiplicative model appropriate? Justify your answers.

Grubb, H. and Mason, A. (2001) Long lead-time forecasting of UK air passengers by Holt–Winters methods with damped trend. *International Journal of Forecasting*, **17**, 71–82.

Do not close the file **airline.sav**. You will need it for Computer Activities 1.7 and 1.8.

When neither a multiplicative model nor an additive model is appropriate for a time series, transformations other than the logarithm function can be tried. Transforming time series using functions such as square roots and other powers is done using **Compute Variable...** from the **Transform** menu.

The use of the **Compute Variable** dialogue box is described in the *Introduction to statistical modelling*.

Computer Activity 1.7 Transforming time series

In Computer Activity 1.6, you found that neither an additive model nor a multiplicative model is appropriate for the time series of monthly numbers of UK airline passengers. Now try the square root transformation, as follows. (Open the SPSS data file **airline.sav**, if it is not already open.)

- ◇ Obtain the **Compute Variable** dialogue box.
- ◇ In the **Target Variable** field, enter the new variable name, **spass** (s for square root, pass for passengers).
- ◇ In the **Numeric Expression** field, type **passengers**(1/2)**.
- ◇ Click on **OK**. The new variable **spass** will appear in the **Data Editor**.

The symbol ****** means ‘raise to the power’, so the expression **passengers**(1/2)** means raise the variable **passengers** to the power $\frac{1}{2}$. There are other ways of calculating square roots, for example, you could use the expression **SQRT(passengers)**. However, the power formulation has been used here as it can be generalized to powers other than $\frac{1}{2}$.

Transform > Compute Variable...

Now obtain a time plot of `spass` by `DATE_`. The default time plot is shown in Figure 1.4.

Use **Analyze > Forecasting > Sequence Charts...** You may need to uncheck the **Natural log transform** box.

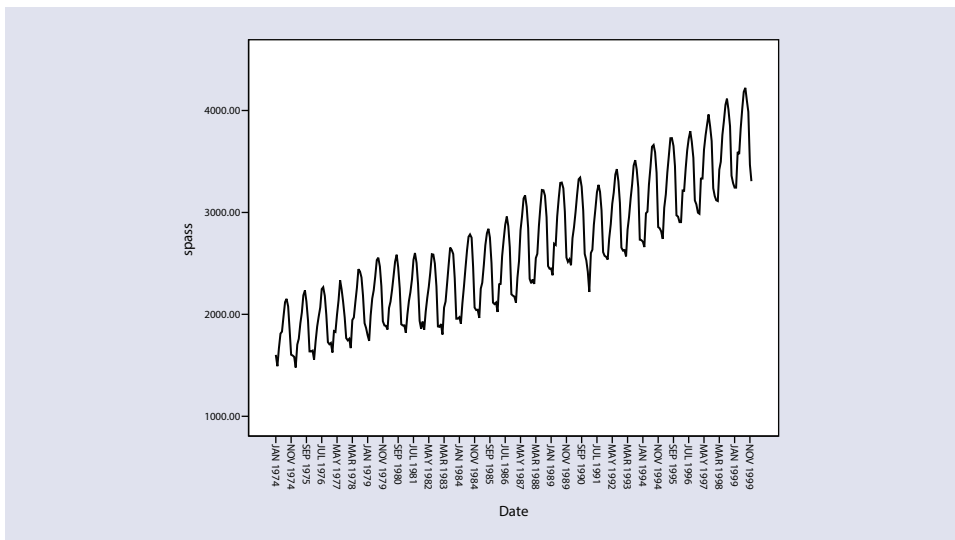


Figure 1.4 Monthly numbers of passengers, square roots

The square root transformation has greatly reduced the change in the size of the seasonal fluctuations, but has perhaps not entirely removed it. The seasonal fluctuations are still a little wider at the end of the series than at the beginning. In Computer Activity 1.8, you will try a different transformation.

Computer Activity 1.8 Another transformation

Open the file **airline.sav** if it is not already open.

- Create a variable named `fpass` to represent the fourth root of the variable `passengers` (the `f` is for fourth). To do this, use the **Compute Variable** dialogue box, and type `passengers**(1/4)` in the **Numeric Expression** field.
- Obtain the time plot of `fpass` by `DATE_`, and comment on the appropriateness of an additive model to represent the transformed time series.

Summary of Chapter 1

In this chapter, you have learned how to enter time series data in SPSS, and how to define dates using the **Define Dates** dialogue box. You have seen that, when appropriate, the seasonal periodicity is determined automatically for use by other SPSS commands. You have learned how to obtain a time plot, how to edit the labelling on the time axis (to make it easier to decide whether or not a time series is seasonal), and how to obtain the time plot for a log transformed series directly. You have also transformed time series data using power transformations, and produced time plots of the transformed series.

Chapter 2

Moving averages

In this chapter, you will learn how to use SPSS to calculate simple centred moving averages. In SPSS, moving averages are calculated using **Create Time Series...** from the **Transform** menu.

Computer Activity 2.1 Smoothing the temperatures

The SPSS data file **alltemps.sav** contains the time series of annual average temperatures recorded in Central England between 1659 and 2004. Open the file now. The file contains three variables: **temperature**, **YEAR_** and **DATE_**. A time plot of this time series is shown in Figure 4.1(a) of *Book 2*; and a simple centred moving average of order 11 was used to smooth the time series. In this activity, you will use SPSS to obtain two further moving averages of this time series.

You will need the file **alltemps.sav** for Computer Activities 2.1 and 2.2.

Create a variable named **ma5**, which contains values of the moving average of order 5 applied to the variable **temperature**, as follows.

- ◇ Choose **Create Time Series...** from the **Transform** menu. The **Create Time Series** dialogue box will open.
- ◇ Enter **temperature** in the **Variable**→**New name** field.
- ◇ Type **ma5** in the **Name** field in the **Name and Function** area to replace the default name.
- ◇ Click on the down arrow in the **Function** field, and select **Centered moving average** from the **Function** drop-down list.
- ◇ Enter 5 in the **Span** field. This specifies the order of the moving average.
- ◇ Click on the **Change** button to register these choices. Notice that the entry in the **Variable**→**New name** field changes.
- ◇ Click on **OK**.

The order of a moving average is called the span by SPSS.

SPSS will indicate in the **Viewer** window that the moving average has been created.

Now look at the **Data View** panel of the **Data Editor**. Notice that the new variable **ma5** has been added to the data set, and that the first two values and the last two values are missing as they cannot be calculated.

Click on the **Variable View** tab. Notice that SPSS has entered a label for **ma5**. Delete this label to reduce clutter. (Do *not* delete the labels for the date variables **YEAR_** and **DATE_**.)

Now use the method just described to create a moving average of order (or span) 31, as follows.

- ◇ Obtain the **Create Time Series** dialogue box.
- ◇ Remove the contents of the **Variable**→**New name** field.
- ◇ Enter the variable **temperature** in the **Variable**→**New name** field.
- ◇ Type **ma31** in the **Name** field to replace the default name.
- ◇ Check that **Centered moving average** is still selected.
- ◇ Enter 31 in the **Span** field.
- ◇ Click on **Change** to register these choices.
- ◇ Click on **OK** and the new variable will be created.
- ◇ Finally, delete the label assigned to the new variable **ma31** in the **Variable View** panel of the **Data Editor**.

Transform > Create Time Series...

You have now created two variables, **ma5** and **ma31**, representing simple centred moving averages of orders 5 and 31, respectively.

Next, obtain a multiple time plot, as follows.

- ◇ Obtain the **Sequence Charts** dialogue box.
- ◇ Enter both **ma5** and **ma31** in the **Variables** field, and **DATE_** in the **Time Axis Labels** field.
- ◇ Leave the **One chart per variable** box unchecked and click on **OK**.

Analyze > Forecasting >
Sequence Charts...

The multiple time plot that will be displayed in the **Viewer** window is shown in Figure 2.1.

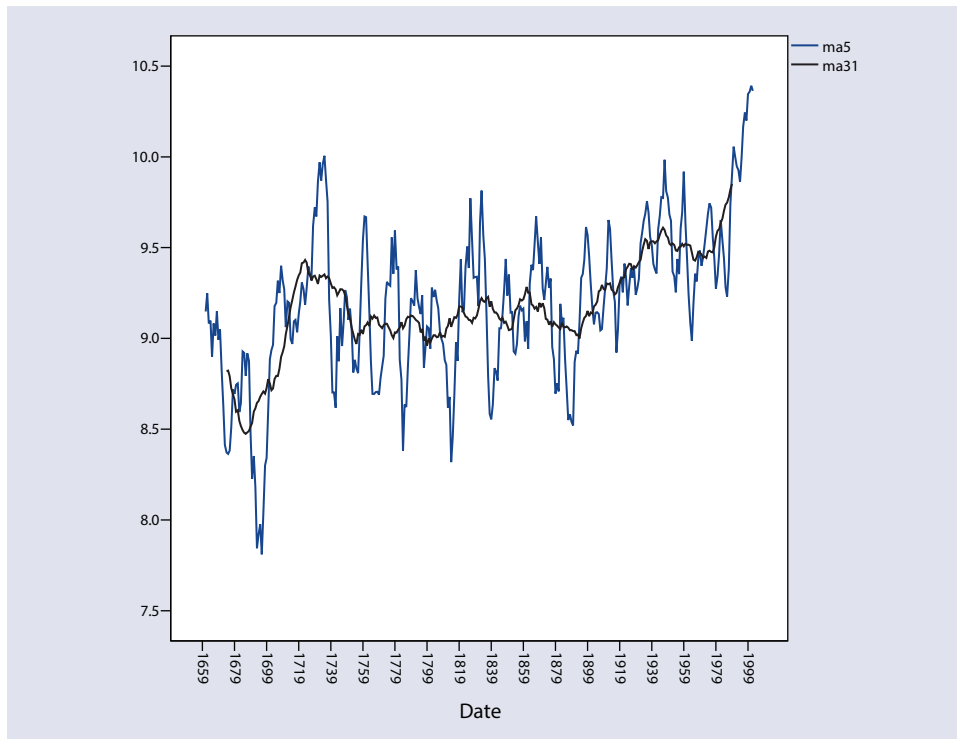


Figure 2.1 Multiple time plot of moving averages of orders 5 and 31

The multiple time plot shows the two moving averages on the same diagram. The time plot of **ma31** is less jagged (and hence smoother) than the time plot of **ma5**. As you saw in Subsection 4.1 of *Book 2*, increasing the order of the moving average increases the amount of smoothing.

Do not close the file **alltemps.sav** as you will need it for Computer Activity 2.2. There is no need to save your work if you are unable to continue directly to Computer Activity 2.2, as the file **alltemps2.sav** contains all the variables from **alltemps.sav**, together with the two moving averages **ma5** and **ma31**.

Computer Activity 2.2 Comparing moving averages

In Computer Activity 2.1, you displayed two moving averages on the same time plot. Sometimes, it is better to display them on separate plots, because superimposed time series can be difficult to interpret. In this activity, you will display the two moving averages on separate plots, then rescale the plots so that they can be compared directly.

You should still have the file **alltemps.sav** open. If not, then open the SPSS data file **alltemps2.sav**.

First, obtain a time plot of `temperature` by `DATE_`. The default time plot is shown in Figure 2.2.

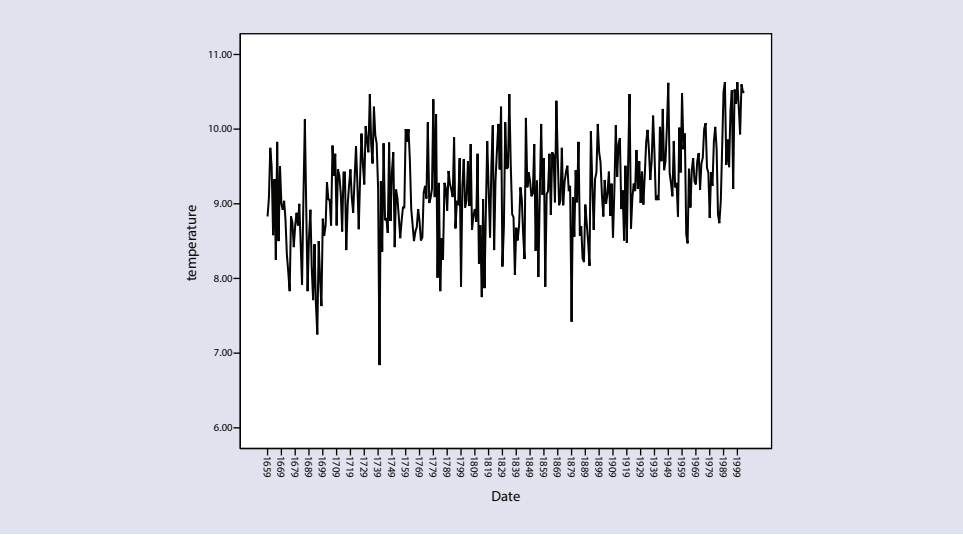


Figure 2.2 Time plot of annual average temperatures

To obtain details of the scale SPSS has used for this time plot, proceed as follows.

- ◇ Double-click on the plot in the **Viewer** window to open the **Chart Editor**.
- ◇ Click on the large **Y** in the **Chart Editor** toolbar. The **Properties** dialogue box will open.
- ◇ Click on the **Scale** tab to bring the **Scale** panel uppermost (if necessary).

Look at the **Range** area of the panel. Notice that the two **Auto** check boxes next to **Minimum** and **Maximum** are checked. This means that SPSS has set the minimum and maximum values of the vertical scale automatically. The minimum and maximum values are given in the **Custom** fields: the minimum is 6 and the maximum is 11. Make a note of these values.

- ◇ Click on **Close** to close the **Properties** dialogue box.
- ◇ Close the **Chart Editor**.

Now obtain the time plot for `ma5`. The default time plot is shown in Figure 2.3.

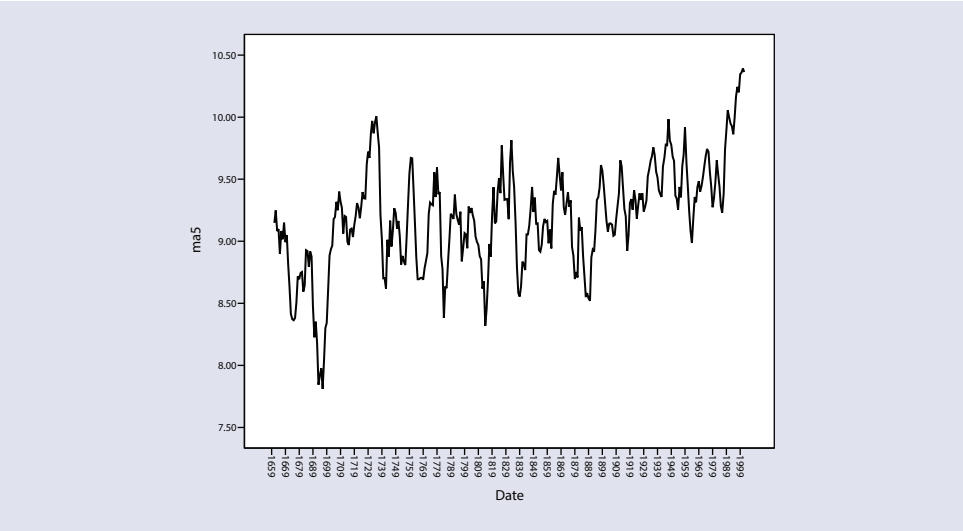


Figure 2.3 Default time plot for `ma5`

Note that the vertical scale on this time plot is different from that on the time plot of the original time series (in Figure 2.2). Change the vertical scale of the time plot for `ma5` to that used for the original time series, as follows.

- ◇ Open the **Chart Editor** by double-clicking on the time plot in the **Viewer** window.
- ◇ Click on the large **Y** in the toolbar of the **Chart Editor** to open the **Properties** dialogue box.
- ◇ In the **Range** area of the **Scale** panel, uncheck the **Auto** check boxes next to **Minimum** and **Maximum** by clicking on them.
- ◇ Replace the value in the **Custom** field for **Minimum** by 6. This sets the minimum of the plotting range to be 6, as for the original time series.
- ◇ Replace the value in the **Custom** field for **Maximum** by 11.
- ◇ Click on **Apply**, then click on **Close**.
- ◇ Close the **Chart Editor**.

The time plot that will be displayed in the **Viewer** window is shown in Figure 2.4(a).

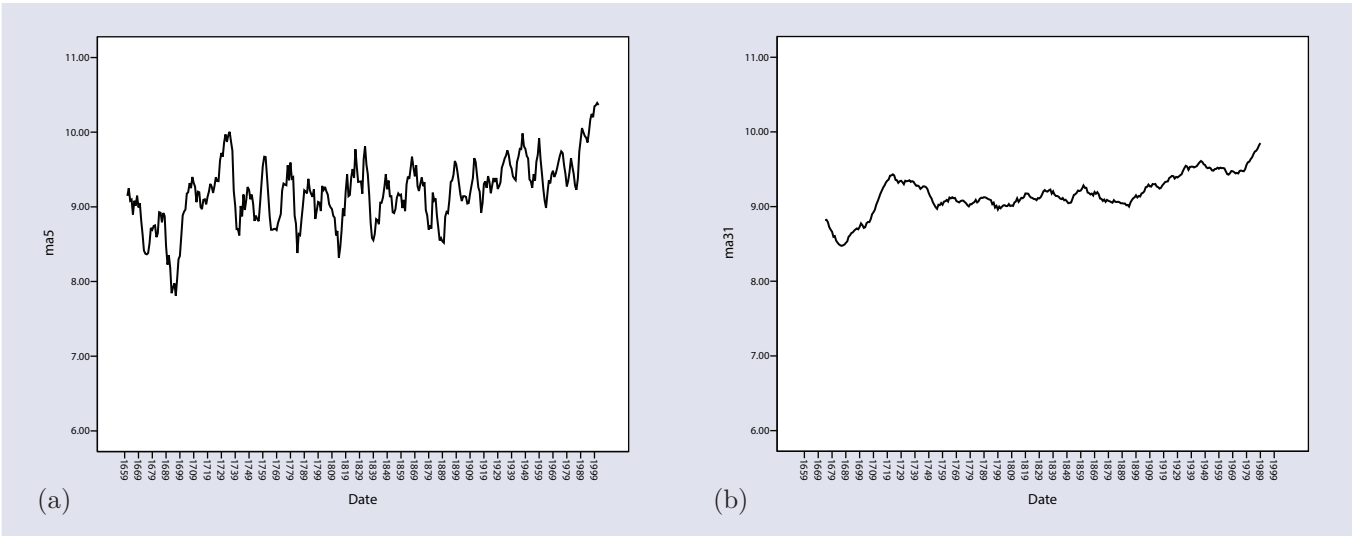


Figure 2.4 Time plots on the same scale as the original series: (a) `ma5` (b) `ma31`

Now use the procedure described above to obtain the time plot for `ma31` drawn on the same scale as the original time series. This will be as shown in Figure 2.4(b).

Now that the scale used for `ma5` and `ma31` is the same as for `temperature`, it is much easier to compare the three plots to evaluate the extent of smoothing. The moving average of order 5 has removed much of the irregular variation, but it is still rather difficult to identify the underlying trend. The moving average of order 31 is much smoother, and perhaps gives a better impression of the trend in temperatures over the past three and a half centuries.

Computer Activity 2.3British Government securities

The SPSS data file `securities.sav` contains four variables: `yield`, `YEAR_`, `MONTH_` and `DATE_`. The data in the variable `yield` are monthly percentage yields on British Government securities, for 21 years between 1950 and 1970. (These years have been chosen for definiteness; the actual dates are approximate.) Open the file now.

- (a) Obtain a time plot of `yield` by `DATE_`. Comment on the main features of the plot. (You may assume that there is no seasonality.)

Chatfield, C. (2004) *The Analysis of Time Series: An Introduction*. Sixth Edition. Chapman & Hall/CRC, London.

- (b) Create a variable named **ma9** containing the simple centred moving average of order 9, and obtain the time plot of **ma9**. Use the time plot to decide in which years, roughly, the peak yields occurred. (You may find it helpful to edit the labelling on the time axis.)
 - (c) What minimum and maximum values for the vertical scale were used to draw the time plot of **ma9**?
 - (d) Create a variable named **ma21** containing the simple centred moving average of order 21, and obtain the time plot of **ma21** drawn on the scale that was used for **ma9** in part (b).
 - (e) Compare the extent of smoothing in the two plots. In your view, which of the two moving averages gives the better summary of the trend? Explain your choice.
-

Summary of Chapter 2

In this chapter, you have learned how to obtain and display simple centred moving averages in SPSS. You have also learned how to alter the scales on which time plots are drawn. You have compared the effect of using moving averages of different orders, by producing time plots drawn using the same vertical scale.

Chapter 3

Estimating the components of a time series

In this chapter, you will learn how to use SPSS to estimate the components of a seasonal time series. This is called seasonal decomposition in SPSS, and is done using **Seasonal Decomposition...** from the **Forecasting** submenu of **Analyze**. In this computer book, the use of this facility assumes that the following conditions are satisfied.

- ◇ The time series has annual seasonality.
- ◇ The time series (or a transformation of it) may be described adequately by an additive model.
- ◇ The time variables and periodicity have been defined in SPSS using **Define Dates...**

In Chapter 1, you examined time plots of the time series of monthly numbers of UK airline passengers. You saw that the time series has annual seasonality, with summer peaks. In Computer Activity 1.8, you found that raising the data to the power $\frac{1}{4}$ results in a series that can be described adequately using an additive model. The use of SPSS to estimate the components of the transformed time series is described in Computer Activities 3.1 to 3.3.

Computer Activity 3.1 *Estimating the seasonal component*

The SPSS data file **airline2.sav** contains all the variables from **airline.sav**, which you used in Computer Activities 1.6, 1.7 and 1.8, together with the additional variable **fpass**, which contains the numbers of monthly passengers raised to the power $\frac{1}{4}$. Open the file now.

First, obtain the time plot of **fpass** by **DATE_**. Figure 3.1 contains a time plot with the labels on the time axis edited to make it easier to check for seasonality.

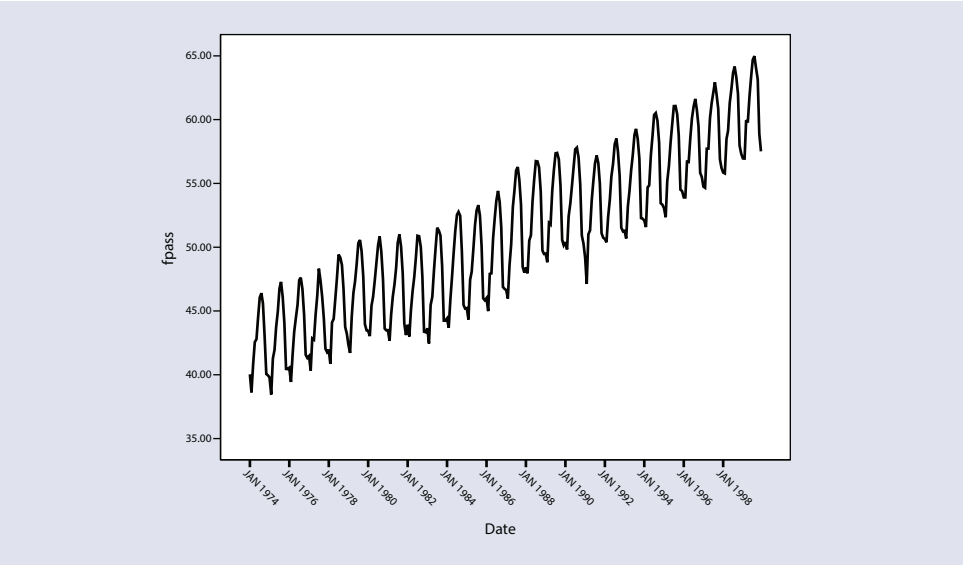


Figure 3.1 Monthly numbers of passengers, fourth roots

There is a roughly linear, increasing trend. There is also seasonal variation. The seasonal fluctuations do not appear to vary in size. The irregular fluctuations are too small compared to the seasonal fluctuations for it to be possible to use this time plot to check that their magnitude remains roughly constant. The assumption that an additive model is valid here will be checked later. At this stage you should simply assume that an additive model is appropriate.

Now obtain estimates of the components of **fpass**, as follows.

- ◇ Choose **Seasonal Decomposition...** from the **Forecasting** submenu of **Analyze**. The **Seasonal Decomposition** dialogue box will open.
- ◇ Enter **fpass** in the **Variable(s)** field.
- ◇ In the **Model Type** area, select **Additive**. This specifies the additive model, which is appropriate for this time series.
- ◇ In the **Moving Average Weight** area, select **Endpoints weighted by .5**.

These are the options needed to specify the moving average that will remove the seasonal component. SPSS recognizes that the period is 12, so it uses the following weighted moving average:

$$SA(t) = \frac{1}{12}(0.5X_{t-6} + X_{t-5} + \cdots + X_t + \cdots + X_{t+5} + 0.5X_{t+6}).$$

- ◇ Click on **OK**. Click again on **OK** at the message that four variables will be added to the data file.

These options are used throughout this computer book.

The estimated seasonal factors will be given in the **Viewer** window in the **Seasonal Factors** output table. (This table is shown in the margin.)

SPSS uses the term ‘period’ for the time points: do not confuse this use of the term ‘period’ with the period of the cycle, which is 12. In this example, periods 1 to 12 are the months January to December. The estimated seasonal factors are positive from May to October, and negative from November to April. The period with the largest positive seasonal component is Period 8 ($\simeq 4.061$), so the seasonal peak is in August. The lowest point is in February, when the seasonal component takes its largest negative value ($\simeq -4.130$). You might like to check for yourself that the estimated seasonal factors sum to zero.

Now look at the **Data View** panel of the **Data Editor**. Four new variables have been created. You will need these in Computer Activity 3.2, so do not close the file **airline2.sav**.

Seasonal Factors	
Series Name: fpass	
Period	Seasonal Factor
1	-3.34887
2	-4.12961
3	-1.46669
4	-.78062
5	1.04079
6	2.41674
7	3.85508
8	4.06110
9	3.19467
10	1.19001
11	-2.83221
12	-3.20039

Computer Activity 3.2 *The seasonally adjusted series*

You should still have the file **airline2.sav** open. If not, then open the SPSS data file **airline3.sav**. This contains the airline passengers time series, plus the four variables created in Computer Activity 3.1.

Look at the variables in the **Data View** panel of the **Data Editor**. The four new variables are called **ERR_1**, **SAS_1**, **SAF_1** and **STC_1**. The most important of these variables is **SAS_1**: this contains the seasonally adjusted series. This is obtained by subtracting the estimated seasonal component (which is in **SAF_1**) from the time series **fpass**.

Click on the **Variable View** tab. Notice that SPSS has given the four new variables very long labels; delete these as they clutter up graphs of the variables. (Do not delete the labels of the date variables **YEAR_**, **MONTH_** and **DATE_**.)

Obtain a time plot of the seasonally adjusted series **SAS_1** by **DATE_**. Figure 3.2 contains a time plot with edited labelling on the time axis.

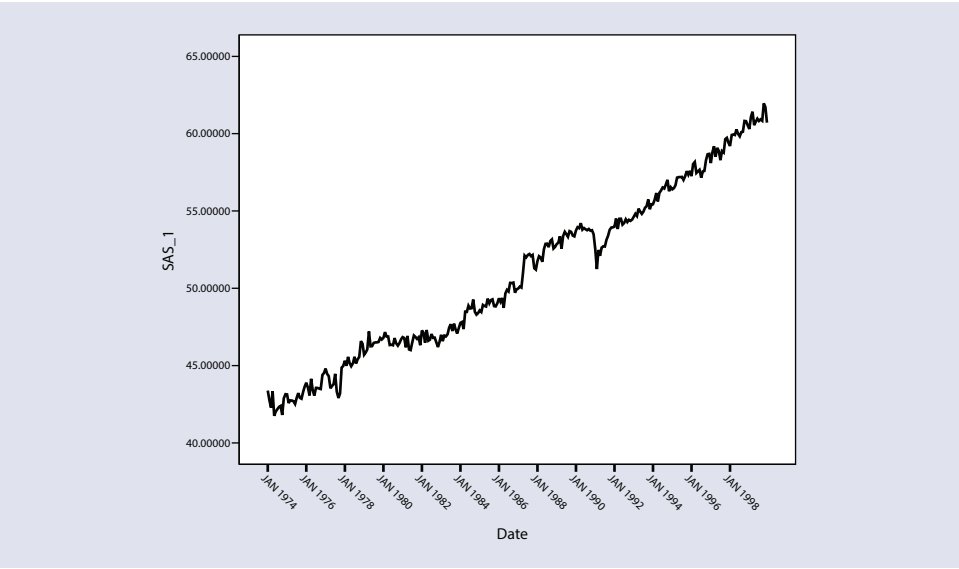


Figure 3.2 Seasonally adjusted time series

Compare this time plot with the time plot in Figure 3.1. The seasonality has been removed from **fpass**, leaving the trend component and the irregular component. Note that it is now apparent that the irregular fluctuations do not vary in size with the level of the series, so the additive model is indeed appropriate.

The variable `STC_1` is a smoothed version of `SAS_1`; it is called the trend-cycle component in SPSS. This name indicates that annual seasonality has been removed, and that the trend and any cycles of period greater than one year remain. Finally, the variable `ERR_1` is an estimate of the irregular component; it is equal to the seasonally adjusted series minus the trend-cycle component.

The variable `STC_1` (and hence also `ERR_1`) is obtained using the same procedure irrespective of the data, but different time series are likely to require different amounts of smoothing. Therefore `STC_1` may not necessarily be smoothed appropriately. Instead of using `STC_1` and `ERR_1`, you will obtain your own estimates of the trend component and the irregular component in Computer Activity 3.3.

Computer Activity 3.3 Completing the decomposition

Open the SPSS data file `airline4.sav`. This is a tidied-up version of `airline3.sav`: the variables `ERR_1`, `SAF_1` and `STC_1` have been deleted; `SAS_1` has been renamed `adjusted`, its long label has been deleted, and the number of decimal places displayed has been reduced from 5 to 2. It is generally a good idea to reduce clutter in this way.

To obtain an estimate of the trend component, smooth the seasonally adjusted series using a simple centred moving average of order 13. Name this moving average `trend`, and obtain a time plot of `trend`.

The method is described in Computer Activity 2.1.

A time plot with edited labelling on the time axis is shown in Figure 3.3(a).

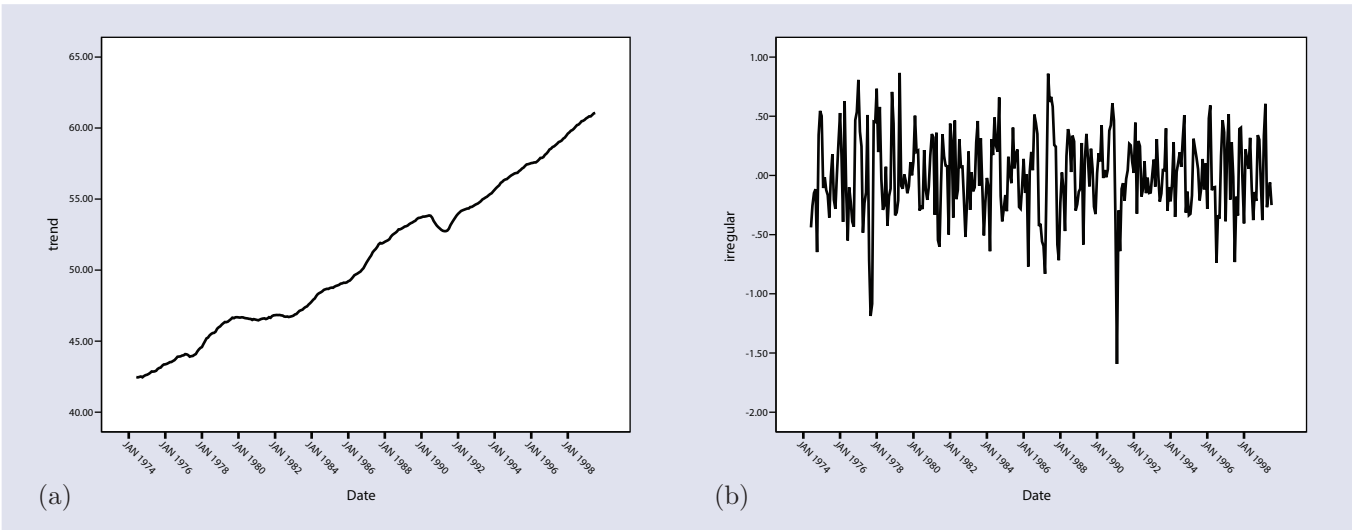


Figure 3.3 Estimated components: (a) `trend` (b) `irregular`

Obtain an estimate of the irregular component by subtracting the trend component from the seasonally adjusted series, as follows.

- ◇ Obtain the **Compute Variable** dialogue box.
- ◇ Type the new variable name `irregular` in the **Target Variable** field.
- ◇ Type the expression `adjusted - trend` in the **Numeric Expression** field.
- ◇ Click on **OK**.

Transform > Compute Variable...

A new variable named `irregular` will be created containing the estimated irregular component of the time series. Obtain a time plot of the irregular component, such as that shown in Figure 3.3(b).

Computer Activity 3.4 will give you some practice at using SPSS to estimate the components of a seasonal time series. This combines many of the SPSS procedures described in Chapters 1 to 3.

Computer Activity 3.4 *Decomposition of temperature data*

The SPSS data file **temperature3.sav** contains the time series of monthly average temperatures (in degrees Celsius) in Central England between January 1970 and December 2004. Open the file now. There are four variables: **temperature**, **YEAR_**, **MONTH_** and **DATE_**. You may assume that an additive model is appropriate for this time series.

- (a) Obtain a time plot of **temperature** by **DATE_**. Edit the plot so that the labels on the time axis are displayed diagonally and occur at two-year intervals. Comment on the appearance of the plot, particularly with respect to trend and seasonality.
 - (b) Obtain the estimated seasonal component of the time series. In what month does the seasonal maximum occur? In what month does the minimum occur?
 - (c) Estimate the trend component by smoothing the seasonally adjusted series with a simple centred moving average of order 35. (The high order is needed because this is a very noisy series.) Obtain a time plot of the trend component, and comment on the plot and what it tells you about the trend.
 - (d) Obtain an estimate of the irregular component of the series, and display it as a time plot. Comment on the relative size of the seasonal variation, the irregular fluctuations, and the change in level over the period 1970–2004. Which component dominates the time series?
-

Summary of Chapter 3

In this chapter, you have learned how to use SPSS to obtain estimates of the components of a time series that can be described adequately by an additive decomposition model. You have learned how to estimate the seasonal factors and obtain the seasonally adjusted series. The use of moving averages to estimate the trend component and the irregular component of a time series has been described.

Chapter 4

Simple exponential smoothing

In this chapter, you will learn how to obtain forecasts using simple exponential smoothing in SPSS. Exponential smoothing methods are carried out using **Create Models...** from the **Forecasting** submenu of **Analyze**. SPSS automatically selects the optimal smoothing parameter and the starting values. Essentially, three steps are involved: selecting the smoothing method; selecting the output required; and specifying the forecasts to be obtained.

In Computer Activity 4.1, the first two of these steps are described; the third will be illustrated in Computer Activity 4.3.

Computer Activity 4.1 Finding the optimal parameter value

The time series of temperatures (in degrees Fahrenheit) of a chemical process was described in *Book 2*. This time series is in the SPSS data file **chemical.sav**. Open the file now.

See *Book 2*, Example 4.3.

There is one variable, **temperature**. A time variable has not been defined as values were recorded every two minutes, whereas SPSS date variables obtained using **Define Dates...** are always at unit intervals.

First obtain a time plot of **temperature**: as there is no date variable, leave the **Time Axis Labels** field of the **Sequence Charts** dialogue box empty. SPSS will produce a time plot with the time axis labelled using the observation numbers 1 to 100, as shown in Figure 4.1.

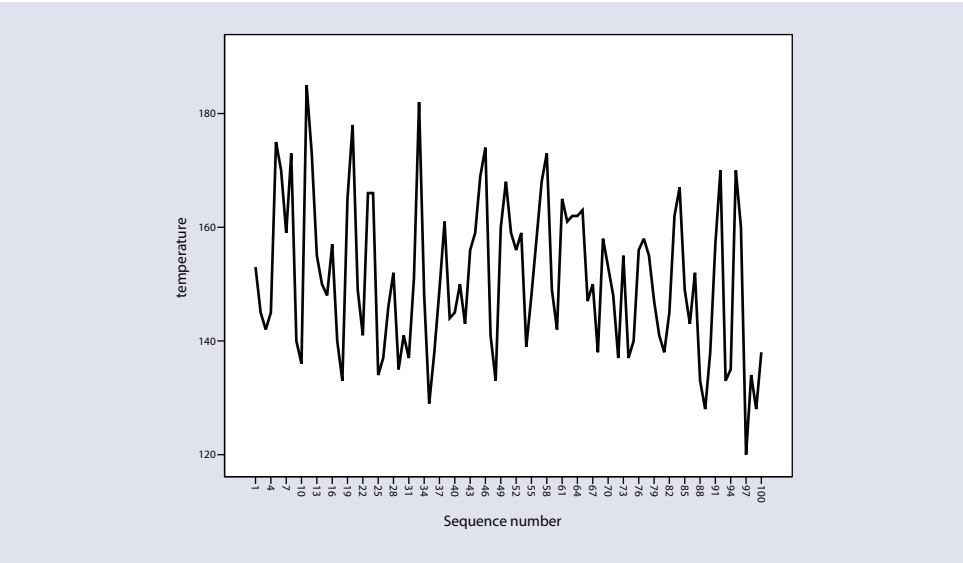


Figure 4.1 Time plot of temperature

The time series is non-seasonal and does not show a marked upward or downward trend. Thus it is reasonable to use the simple exponential smoothing method to obtain forecasts.

In this activity, you will obtain the optimal smoothing parameter α for these data. Instead of the *SSE* (the Sum of Squared Errors), SPSS uses a related quantity to assess the accuracy of the 1-step ahead forecasts. This related quantity is the Root Mean Squared Error, or *RMSE*.

The relationships between the *RMSE* and the *SSE* are:

$$RMSE = \sqrt{\frac{SSE}{n-k}}, \quad SSE = (n-k) \times RMSE^2,$$

where n is the number of observed values (so here, $n = 100$) and k is the number of smoothing parameters. For simple exponential smoothing there is just one smoothing parameter, α , so $k = 1$. Note that for a given exponential smoothing method, minimizing the *SSE* is the same as minimizing the *RMSE*.

First, enter the variable name and select the smoothing method, as follows.

- ◇ Choose **Create Models...** from the **Forecasting** submenu of **Analyze**. The **Time Series Modeler** dialogue box will open with a prompt because a time variable has not been defined.
- ◇ Click on **OK** to dismiss the prompt: the data will be treated as unspecified time periods consecutively labelled as 1, 2, 3, ..., 100.
- ◇ The **Time Series Modeler** dialogue box contains six tabbed panels. If necessary, click on the **Variables** tab at the top left-hand corner of the **Time Series Modeler** dialogue box so that the **Variables** panel is uppermost.
- ◇ Enter **temperature** in the **Dependent Variables** field. Leave the **Independent Variables** field empty.
- ◇ Click on the down arrow on the right of the **Method** box and select **Exponential Smoothing** from the **Method** drop-down list.
- ◇ Click on the **Criteria...** button. The **Time Series Modeler: Exponential Smoothing Criteria** dialogue box will open.
- ◇ In the **Model Type** area, select **Simple** under **Nonseasonal**. In the **Dependent Variable Transformation** area, make sure that **None** is selected.
- ◇ Click on **Continue**. Note that the selected model is described as **Model Type: Simple nonseasonal** below the **Method** box.

Next, select the output required. SPSS produces a large amount of output; the following steps are to select exactly what is required and no more.

- ◇ Click on the **Statistics** tab to display the **Statistics** panel.
- ◇ Make sure the **Display fit measures, Ljung-Box statistic, and number of outliers by model** box is checked.
- ◇ In the **Fit Measures** area, deselect **Stationary R square** and select **Root mean square error**.
- ◇ Uncheck any checked boxes in the **Statistics for Comparing Models** area. In the **Statistics for Individual Models** area, select **Parameter estimates** (and nothing else).
- ◇ Now click on the **Plots** tab.
- ◇ Leave unchecked all boxes within the **Plots for Comparing Models** area. In the **Plots for Individual Models** area, select **Series**. In the **Each Plot Displays** sub-area, select **Observed values** and **Fit values** and deselect **Forecasts**, leaving the other boxes unchecked.

Do not click on **OK** at this stage. If you do, SPSS will run with an incomplete set-up. If this happens, just start again.

The set-up is now complete.

- ◇ Click on **OK** and the simple exponential smoothing forecasts will be obtained.

The following three output tables will be displayed in the **Viewer** window.

Model Description

			Model Type
Model ID	temperature	Model_1	Simple

Model Statistics

Model	Number of Predictors	Model Fit statistics	Ljung-Box Q(18)			Number of Outliers
		RMSE	Statistics	DF	Sig.	
temperature-Model_1	0	13.565	41.541	17	.001	0

Exponential Smoothing Model Parameters

Model			Estimate	SE	t	Sig.
temperature-Model_1	No Transformation	Alpha (Level)	.056	.040	1.411	.161

The **Model Description** table describes the model, which has been labelled **temperature-Model_1**. The third column of the **Model Statistics** table gives the value of the *RMSE*, which is 13.565. The second column of the **Exponential Smoothing Model Parameters** table gives the optimal value of the smoothing parameter, $\alpha = 0.056$. You should ignore all other entries in these tables.

Note that the optimal value of α is close to zero, so recent observations have little influence on the 1-step ahead forecasts. Since $n = 100$ and $k = 1$, the *SSE* is

$$SSE = (n - 1) \times RMSE^2 = (100 - 1) \times 13.565^2 \simeq 18\,217.$$

Below the output tables is a multiple time series plot, similar to that shown in Figure 4.2, showing the observed values and the 1-step ahead forecasts obtained by simple exponential smoothing with the optimal value $\alpha = 0.056$.

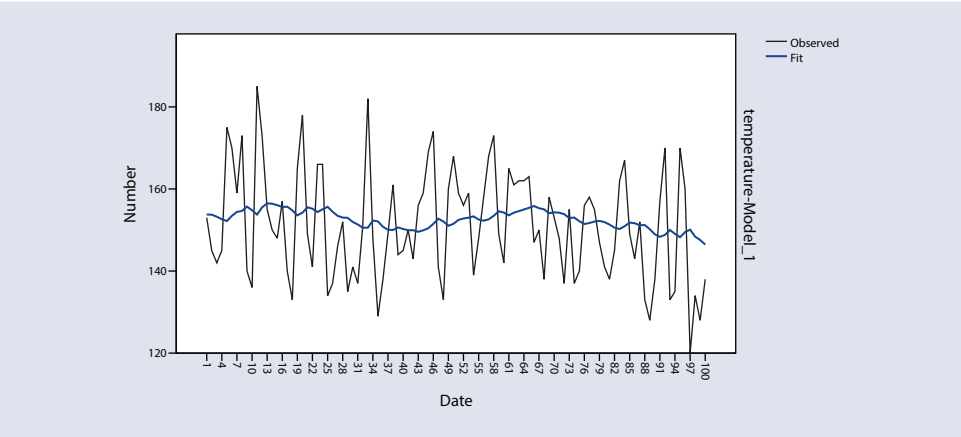


Figure 4.2 Observed and forecasted values

The forecasted values (which SPSS calls Fit values) are much smoother than the observed values. This reflects the fact that the optimal value of the smoothing parameter α is close to zero.

Computer Activity 4.2 Precipitation in Morocco

This activity will give you some practice at carrying out simple exponential smoothing in SPSS.

The SPSS data file **Taourirt.sav** contains data on the annual precipitation (in mm) in Taourirt in Morocco, for 31 successive years starting in 1967/68. Open the file now. There is a single variable, **precipitation**.

Chigneur, M. (1999) Déséquilibre du marché du travail, niveaux des revenus et émigration: Cas de l’Oriental du Maroc. *International Colloquium on the Maghreb and Europe at the Dawn of the Twenty-first Century*. Quadi Ayyad University, Marrakech.

- (a) Obtain a time plot of **precipitation**, and discuss the appropriateness of using simple exponential smoothing to obtain forecasts.
- (b) Using simple exponential smoothing, obtain the optimal value of the smoothing parameter α and the *RMSE*. Calculate the *SSE* (to two decimal places).
- (c) Display the observed and forecasted values in a multiple time plot. Comment briefly on the accuracy of the forecasts.

Computer Activity 4.3 *Forecasting the temperature of a chemical process*

In this activity, you will learn how to use SPSS to forecast future values of the chemical process data that you used in Computer Activity 4.1, beyond the last observation at time point 100.

In *Book 2*, only 1-step ahead forecasts were discussed. However, SPSS will calculate forecasts for any required number of steps ahead. While it is inadvisable to use statistical forecasting methods to predict future observations a very long time ahead, forecasts more than a single step ahead are often needed. In this activity, you will obtain simple exponential smoothing forecasts up to the 105th observation point, and learn how to save the forecasts and the forecast errors.

The data are in the SPSS data file **chemical.sav**. Open this file now; if it is already open, use **Windows** on the main toolbar to make it active.

A quick way to open the data file is to use **Recently Used Data** from the **File** menu.

Some of this activity repeats Computer Activity 4.1. However there are some additional steps and selections. The whole procedure is thus described in detail, for completeness.

- ◇ Choose **Create Models...** from the **Forecasting** submenu of **Analyze**. The **Time Series Modeler** dialogue box will open, together with the prompt because a time variable has not been defined.
- ◇ Click on **OK** to dismiss the prompt.
- ◇ Click on **Reset**.
- ◇ Click on the **Variables** tab so that the **Variables** panel is uppermost.
- ◇ Enter **temperature** in the **Dependent Variables** field. Leave the **Independent Variables** field empty.
- ◇ Select **Exponential Smoothing** from the **Method** drop-down list. Check that the selected model is described as **Model Type: Simple nonseasonal** below the **Method** box.
- ◇ Click on the **Statistics** tab.
- ◇ Make sure, as before, that the following three boxes are checked: **Display fit measures**, **Ljung–Box statistic**, and **number of outliers by model**; **Root mean square error**; and **Parameter estimates**. Uncheck any other checked boxes.
- ◇ In addition, check the **Display forecasts** box.
- ◇ Now click on the **Plots** tab. Leave unchecked all the boxes in the **Plots for Comparing Models** area. In the **Plots for Individual Models** area, select **Series**. In the **Each Plot Displays** sub-area, select **Observed values**, **Forecasts** and **Fit values**.

If it is not, click on **Criteria** and make the selections required.

The forecasts and the 1-step ahead prediction errors can be added to the data file using the **Save** panel, as follows:

- ◇ Click on the **Save** tab.
- ◇ Under **Variables** in the **Save Variables** area, check the boxes in the **Save** column corresponding to **Predicted Values** and **Noise Residuals**.

Finally, specify the range of future time points at which forecasts are required using the **Options** panel.

- Click on the **Options** tab.
- Select **First case after end of estimation period through a specified date** in the **Forecast Period** area.
- Type 105 in the **Observation** field under **Date**. (105 is the last time point for which a forecast is required.)

The set-up is now complete.

- Click on **OK**.

In the **Viewer** window, the three output tables obtained in Computer Activity 4.1 will be displayed, together with the following output table.

Forecast						
Model		101	102	103	104	105
temperature-Model_1	Forecast	146	146	146	146	146
	UCL	173	173	173	173	173
	LCL	119	119	119	119	119

For each model, forecasts start after the last non-missing in the range of the requested estimation period, and end at the last period for which non-missing values of all the predictors are available or at the end date of the requested forecast period, whichever is earlier.

The first row of the **Forecast** table gives the values forecasted for time periods 101 to 105. Since simple exponential smoothing assumes that there is no trend, these forecasts are all the same and in this case are equal to 146.

The observed values, together with the forecasted values, are shown in a multiple time plot similar to that in Figure 4.3.

The last two rows of the table give 95% upper and lower prediction limits for the forecasts. These will be discussed briefly in *Chapter 6*.

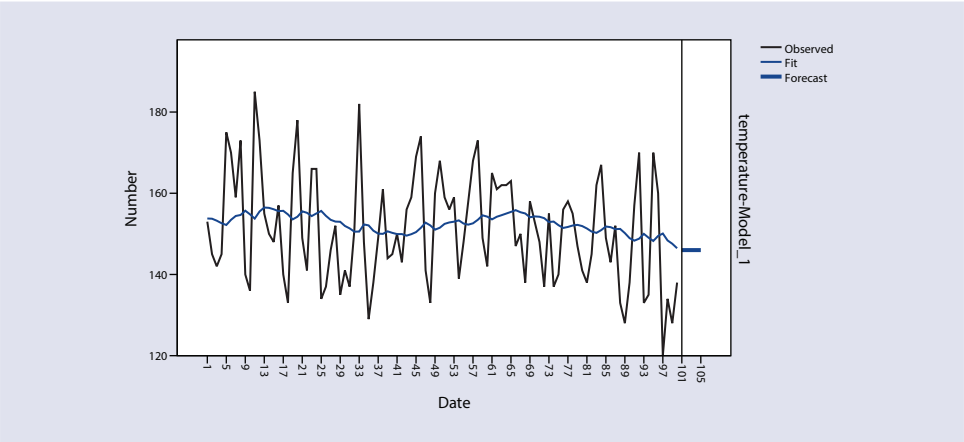


Figure 4.3 Observed and forecasted values

The vertical bar at time point 101 separates the 1-step ahead forecasts based on the data values (which SPSS calls Fit values), from the values forecasted for time points 101 to 105.

Finally, look at the **Data View** panel of the **Data Editor**. Two new variables have been created, each with very long names. The variable **Predicted_temperature_Model_1** contains the forecasts, and the variable **NResidual_temperature_Model_1** contains the 1-step ahead prediction errors, which are the differences between the observed values and the 1-step ahead forecasts at time points 1 to 100. Scroll down to the bottom of the file and notice that the errors are calculated to time point 100, whereas the forecasts carry on to time point 105. The observations and errors at time points 101 to 105 are not available, and are replaced by dots indicating missing values.

It is good practice to tidy up the file by renaming the variables with shorter names (such as **forecast** and **error**) and deleting the variable labels in the **Variable View** panel of the **Data Editor**. The tidied-up file has been saved as **chemical2.sav**.

Computer Activity 4.4 Forecasting precipitation in Taourirt

In this activity, you will obtain forecasts for the Moroccan precipitation data. The data are in the file **Taourirt.sav**. Open this file now.

- (a) Forecast the precipitation for time points 32 to 35 (corresponding to the years 1998/9 to 2001/2) using simple exponential smoothing. What are the forecasted values for the years 1998/9 to 2001/2?
- (b) Display the observed and forecasted values in a multiple time plot. Comment briefly on the forecasts for the years 1998/9 to 2001/2.

Summary of Chapter 4

In this chapter, you have learned how to obtain forecasts in SPSS using simple exponential smoothing. Finding the optimal value of the smoothing parameter, and the corresponding values of the *RMSE* and the *SSE*, have been described.

Chapter 5 Holt's and Holt–Winters exponential smoothing

In this chapter, you will learn how to use SPSS to obtain forecasts using Holt's and Holt–Winters exponential smoothing. These methods are carried out using **Create Models...** from the **Forecasting** submenu of **Analyze**. Holt's method is discussed in Computer Activity 5.1.

Computer Activity 5.1 Airline passenger numbers

The SPSS data file **airline5.sav** contains data on the annual numbers of airline passengers in the UK from 1949 to 1999. Open the file now. There are four variables: **passengers**, **fpass**, **YEAR_** and **DATE_**. The variable **fpass** contains the fourth roots of the numbers of passengers.

Forecasts are required for the numbers of airline passengers for the years 2000 to 2004. These forecasts will be obtained using the time series of fourth roots of the numbers of passengers. Before applying a smoothing method, you should check that it is appropriate by looking at a time plot of the data.

Obtain a time plot of **fpass** by **DATE_**, such as that shown in Figure 5.1.

You explored the time series of monthly numbers of passengers in Chapters 1 and 3.

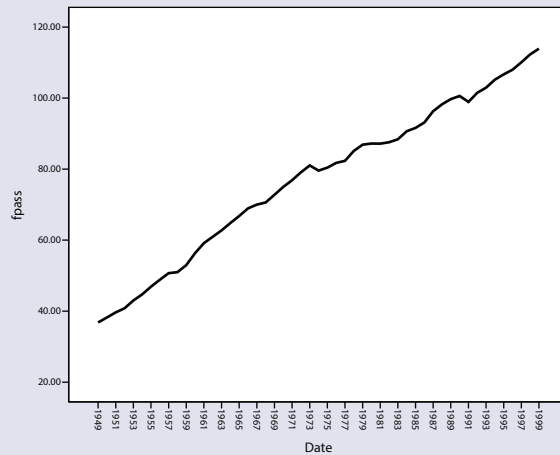


Figure 5.1 Annual numbers of passengers, fourth roots

The time series shows a linear increasing trend. The data are annual, so the time series is not seasonal. Hence Holt's exponential smoothing method is appropriate.

Obtain forecasts using Holt's method, as follows.

- ◇ Obtain the **Time Series Modeler** dialogue box.
- ◇ Click on the **Variables** tab.
- ◇ Enter **fpass** in the **Dependent Variables** field. Leave the **Independent Variables** field empty.
- ◇ Select **Exponential Smoothing** from the **Method** drop-down list.
- ◇ Click on the **Criteria...** button. The **Time Series Modeler: Exponential Smoothing Criteria** dialogue box will open.
- ◇ In the **Model Type** area, select **Holt's linear trend** under **Nonseasonal**. In the **Dependent Variable Transformation** area, make sure that **None** is selected.
- ◇ Click on **Continue**. Note that the selected model is described as **Model Type: Holt's linear trend** below the **Method** box.

Next, select the output required. The method is exactly the same as for simple exponential smoothing, but is described in full for completeness.

- ◇ Click on the **Statistics** tab.
- ◇ Make sure the **Display fit measures**, **Ljung–Box statistic**, and **number of outliers by model** box is checked.
- ◇ In the **Fit Measures** area, deselect **Stationary R square** and select **Root mean square error**.
- ◇ Uncheck any checked boxes in the **Statistics for Comparing Models** area.
- ◇ In the **Statistics for Individual Models** area, select **Parameter estimates** (and nothing else).
- ◇ Select **Display forecasts**.
- ◇ Now click on the **Plots** tab.
- ◇ Leave unchecked all boxes within the **Plots for Comparing Models** area.
- ◇ In the **Plots for Individual Models** area, select **Series**.
- ◇ In the **Each Plot Displays** sub-area, select **Observed values**, **Forecasts** and **Fit values**, leaving the other boxes unchecked.

Analyze > Forecasting > Create Models... This time there is no prompt, because the time variable has been defined.

The forecasts and the 1-step ahead prediction errors can be added to the data file as follows.

- Click on the **Save** tab.
- Under **Variables** in the **Save Variables** area, select **Predicted Values** and **Noise Residuals** by checking their boxes in the **Save** column.

Finally, specify the range of future time points at which forecasts are required.

- Click on the **Options** tab.
- Select **First case after end of estimation period through a specified date** in the **Forecast Period** area.
- Under **Date**, type 2004 in the **Year** field. (2004 is the last year for which a forecast is required.) Leave all other settings unchanged.

In Chapter 4, the **Year** field was called **Observation** because SPSS date variables were not used.

The set-up is now complete.

- Click on **OK** to obtain Holt’s exponential smoothing forecasts.

The output is similar to that obtained with simple exponential smoothing. The **Exponential Smoothing Model Parameters** output table is shown below.

Exponential Smoothing Model Parameters						
Model			Estimate	SE	t	Sig.
fpass-Model_1	No Transformation	Alpha (Level)	1.000	.144	6.953	.000
		Gamma (Trend)	.001	.027	.035	.972

The optimal values of the smoothing parameters are $\alpha = 1.000$ for the level and $\gamma = 0.001$ for the trend. You should ignore the other entries in this table. Since $\alpha = 1$, past values are not used in estimating the level, while the value $\gamma = 0.001$ means that recent values have hardly any influence on the trend, which remains roughly constant throughout the time series.

The **Model Statistics** output table shows that the *RMSE* is 0.975. Since there are $n = 51$ observations (as you can check by scrolling down the **Data View** panel of the **Data Editor**) and $k = 2$ smoothing parameters, the *SSE* is

$$SSE = (51 - 2) \times 0.975^2 \simeq 46.6.$$

The **Forecast** output table is shown below.

Forecast						
Model		2000	2001	2002	2003	2004
fpass-Model_1	Forecast	115.42	116.91	118.39	119.88	121.37
	UCL	117.38	119.68	121.79	123.81	125.76
	LCL	113.46	114.13	115.00	115.96	116.98

For each model, forecasts start after the last non-missing in the range of the requested estimation period, and end at the last period for which non-missing values of all the predictors are available or at the end date of the requested forecast period, whichever is earlier.

The table gives the forecasts for 2000 to 2004. Note that, unlike the forecasts for simple exponential smoothing, these increase over time, in line with the trend in Holt’s exponential smoothing method. The observed values and the forecasts are displayed in a multiple time plot similar to that shown in Figure 5.2.

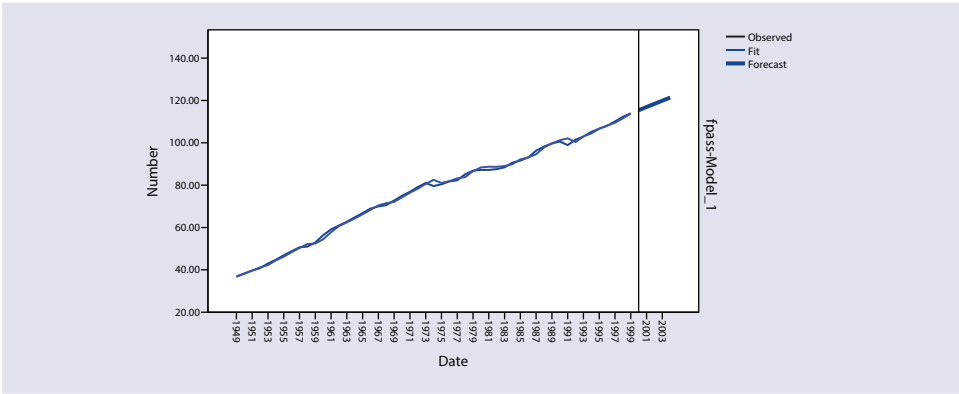


Figure 5.2 Observed and forecasted values of fpass

The vertical bar at year 2000 separates the 1-step ahead forecasts based on the data values (which SPSS calls Fit values), from the values forecasted for the years 2000 to 2004.

If required, the forecasts can be transformed back to the original scale by raising them to the power 4. (This is the inverse operation of taking fourth roots.) Thus, for example, the forecasted number of airline passengers for 2004 is

$$121.37^4 \simeq 217\,000\,000$$

to the nearest hundred thousand.

Finally, look at the **Data View** panel of the **Data Editor**. As for simple exponential smoothing, two new variables have been created. The variable **Predicted_fpass_Model_1** contains the forecasts, and the variable **NResidual_fpass_Model_1** contains the 1-step ahead prediction errors, which are the differences between the observed values and the 1-step ahead forecasts at years 1949 to 1999. A tidied-up version of the file, with the new variables renamed **forecast** and **error** and their variable labels deleted, has been saved in **airline6.sav**.

Computer Activity 5.2 UK index of production

The quarterly UK index of production measures the activity of industries in the manufacturing, extractive and energy supply sector. The series is seasonally adjusted. Values of the index from the first quarter of 1990 to the first quarter of 2005 are in the SPSS data file **production.sav**. Open the file now. There are four variables: **index**, **YEAR_**, **QUARTER_** and **DATE_**.

- Obtain a time plot of **index** by **DATE_**. Comment on the suitability of Holt's exponential smoothing method for forecasting the production index in 2005.
- Forecast the production index for the last three quarters of 2005, using Holt's exponential smoothing method. Note that SPSS will recognize that these data are quarterly. (Under **Date** in the **Forecast Period** area of the **Options** panel of the **Time Series Modeler** dialogue box, you should type 2005 in the **Year** field, and 4 in the **Quarter** field.)

Obtain the optimal values of the smoothing parameters and the *RMSE* for these parameter values. Calculate the *SSE*.

- Obtain a multiple time plot of the observed and forecasted values. Briefly summarize your results.

These data were obtained from the National Statistics website <http://www.statistics.gov.uk> in February 2005.

Crown copyright material is reproduced with the permission of the Controller of HMSO.

Computer Activity 5.3 Quarterly *Clostridium difficile* reports

The SPSS data file **infections2.sav** contains quarterly data on the numbers of reports from England and Wales of several infectious agents between the third quarter of 1996 and the second quarter of 2003. One of these infections is *Clostridium difficile*. In this activity, you will use Holt–Winters exponential smoothing to obtain forecasts for each quarter up to the second quarter of 2004.

Open the file now. There are several variables in the data file: those of interest for this activity are **clostridium**, **YEAR_**, **QUARTER_** and **DATE_**.

Begin by checking that the proposed method is appropriate: obtain a time plot of **clostridium** by **DATE_**, such as that shown in Figure 5.3.

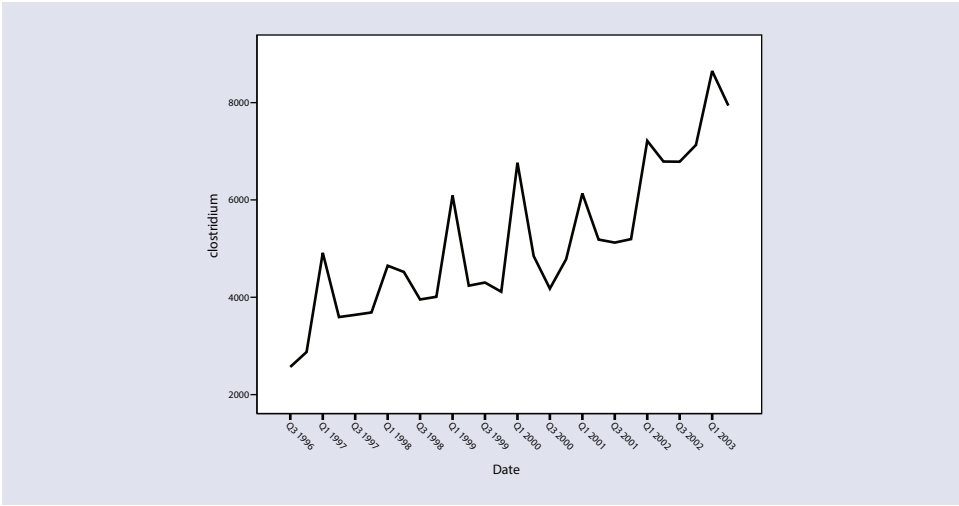


Figure 5.3 Reports of *Clostridium difficile*

The time plot in Figure 5.3 shows an upward trend, and there are seasonal peaks. In *Book 2*, it was suggested that an additive model may be suitable for the time series of weekly numbers of infections. In this activity, it will be assumed that an additive model is also suitable for the quarterly data. Hence the Holt–Winters exponential smoothing method is appropriate.

See Example 2.2 of *Book 2*.

Now specify the model, as follows.

- ◇ Obtain the **Time Series Modeler** dialogue box.
- ◇ Click on the **Variables** tab.
- ◇ Enter **clostridium** in the **Dependent Variables** field. Leave the **Independent Variables** field empty.
- ◇ Select **Exponential Smoothing** from the **Method** drop-down list.
- ◇ Click on the **Criteria...** button. The **Time Series Modeler: Exponential Smoothing Criteria** dialogue box will open.
- ◇ In the **Model Type** area, select **Winter's additive** under **Seasonal**. In the **Dependent Variable Transformation** area, make sure that **None** is selected.
- ◇ Click on **Continue**. Note that the selected model is described as **Model Type: Winter's additive** below the **Method** box.

The remaining settings, which are the same as for simple exponential smoothing and Holt's exponential smoothing, are selected as follows.

- Click on the **Statistics** tab.
- Make sure that the **Display fit measures, Ljung–Box statistic, and number of outliers by model** box is checked. Select **Root mean square error, Parameter estimates** and **Display forecasts**, and deselect any other options if their boxes are checked.
- Now click on the **Plots** tab.
- Leave unchecked all boxes within the **Plots for Comparing Models** area. In the **Plots for Individual Models** area, select **Series**. In the **Each Plot Displays** sub-area, select **Observed values, Forecasts** and **Fit values**, leaving the other boxes unchecked.

Finally, specify the range of future time points at which forecasts are required.

- Click on the **Options** tab.
- Select **First case after end of estimation period through a specified date** in the **Forecast Period** area.
- Under **Date**, type 2004 in the **Year** field and 2 in the **Quarter** field. (The second quarter of 2004 is the last time period for which a forecast is required.) Leave all other settings unchanged.

The set-up is now complete.

- Click on **OK** to obtain the Holt–Winters forecasts.

The output is similar to that obtained with the other exponential smoothing methods. The **Exponential Smoothing Model Parameters** output table is shown below.

In this activity, the forecasts and 1-step ahead forecast errors need not be added to the data file, so you do not need to use the **Save** panel.

Exponential Smoothing Model Parameters						
Model			Estimate	SE	t	Sig.
clostridium-Model_1	No Transformation	Alpha (Level)	.598	.206	2.912	.007
		Gamma (Trend)	3.798E-006	.103	3.671E-005	1.000
		Delta (Season)	3.003E-006	.354	8.485E-006	1.000

The optimal values of the smoothing parameters are $\alpha = 0.598$ for the level, $\gamma = 3.798\text{E-}006$ for the trend and $\delta = 3.003\text{E-}006$ for the seasonal component. You should ignore the other entries in this table. Since $\alpha = 0.598$, recent values are moderately influential in estimating the level. The values of γ and δ are very close to zero, so recent values have hardly any influence on the trend or the seasonal component. This means that the trend and seasonal components are roughly constant throughout the time series.

The **Model Statistics** output table shows that the *RMSE* is 427.651. Since there are $n = 28$ observations (as you can check by scrolling down the **Data View** panel of the **Data Editor**) and $k = 3$ smoothing parameters, the *SSE* is

$$SSE = (28 - 3) \times 427.651^2 \simeq 4\,572\,130.$$

The **Forecast** output table is shown below.

Forecast					
Model		Q3 2003	Q4 2003	Q1 2004	Q2 2004
clostridium-Model_1	Forecast	7488	7665	9468	8425
	UCL	8369	8692	10622	9694
	LCL	6607	6639	8314	7156

For each model, forecasts start after the last non-missing in the range of the requested estimation period, and end at the last period for which non-missing values of all the predictors are available or at the end date of the requested forecast period, whichever is earlier.

The table gives the forecasts for each quarter from the third quarter of 2003 to the second quarter of 2004. The forecast for the second quarter of 2004 is 8425. The observed values and the forecasts are displayed in a multiple time plot similar to that shown in Figure 5.4.

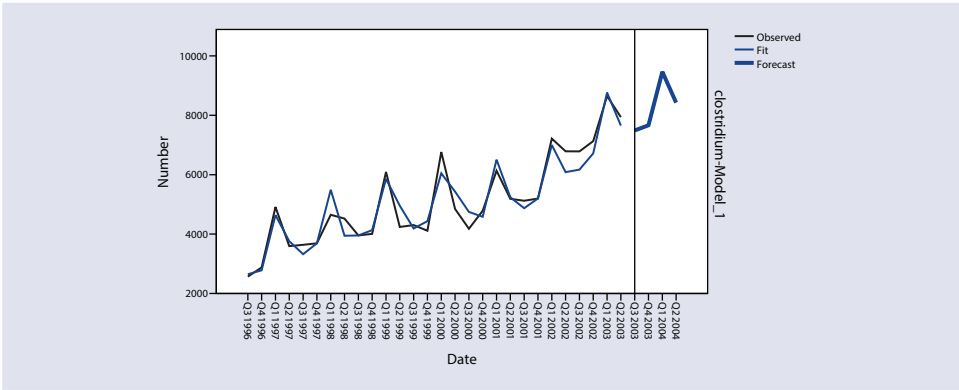


Figure 5.4 Observed and forecasted monthly infections

Notice that both the trend and the seasonal variation have been taken into account in making the forecasts for the last four quarters.

Computer Activity 5.4 Monthly demand for electricity

This activity will give you some practice at using the Holt–Winters exponential smoothing method, this time with monthly data.

The SPSS data file **electricity.sav** contains data on the monthly demand for electricity (in megawatt-hours) at a utility in California, between January 1965 and December 1989. Open the file now. There are five variables: **demand**, **YEAR_**, **MONTH_**, **DATE_** and **sademand**. The variable **sademand** will not be used in this activity.

Delurgio, S.A. (1998) *Forecasting Principles and Applications*. McGraw-Hill, Singapore.

In this activity, you will use the Holt–Winters method to obtain forecasts for the year 1990, based on the data up to 1989.

- (a) Obtain a time plot of **demand** by **DATE_**, and comment on the appropriateness of an additive model for these data.
- (b) Create a variable named **logdemand**, containing the natural logarithm of the variable **demand**. Obtain a time plot of **logdemand** by **DATE_**. Is an additive model appropriate for these data?
- (c) Use the Holt–Winters method to obtain forecasts for **logdemand** in the twelve months from January to December 1990. Write down the optimal values of the parameters, and the corresponding *RMSE*. Calculate the *SSE*.
- (d) Obtain a multiple time plot of the observed and forecasted values of **logdemand** from January 1965 to December 1990. Comment briefly on this plot. Use the time plot or the output file to say in which month of 1990 the forecasted demand is lowest and in which month it is highest.

Use **Transform > Compute Variable...**
(See Computer Activity 1.7.)

Summary of Chapter 5

In this chapter, the application of Holt’s and Holt–Winters exponential smoothing in SPSS has been described. You have learned how to select an appropriate forecasting method in SPSS and how to obtain the optimal combination of smoothing parameters.

Chapter 6

Correlograms and forecasting

In Chapters 4 and 5, you learned how to obtain forecasts in SPSS using the exponential smoothing methods described in Sections 6 and 7 of *Book 2*. In this chapter, you will learn how to apply the techniques for evaluating the performance of these forecasting methods that are described in Section 9 of *Book 2*; these include the correlogram and prediction intervals. The method for calculating a prediction interval for a 1-step ahead forecast depends on the assumption that the forecast errors are white noise — that is, they are normally distributed with mean zero and constant variance, and the autocorrelations at lags $k \geq 1$ are all zero. You will learn how to use SPSS to check that the white noise assumption is reasonable: obtaining the correlogram and carrying out the Ljung–Box test are described; and a graphical method for checking the normality assumption is introduced. The chapter ends with a complete forecasting exercise, which brings together many of the SPSS skills that you have acquired so far.

The term ‘white noise’ is defined in Subsection 9.3 of *Book 2*.

Computer Activity 6.1 Correlogram for airline forecast errors

In Computer Activity 5.1, you applied Holt’s exponential smoothing method to the fourth roots of the annual numbers of airline passengers in the UK between 1949 and 1999. The 1-step ahead forecasts and forecast errors from this analysis have been saved in the SPSS data file **airline6.sav**. Open the file now. The forecast errors are stored in the variable **error**.

Correlograms are produced in SPSS using **Autocorrelations...** from the **Forecasting** submenu of **Analyze**. Obtain the correlogram for lags 1 to 20 for the forecast errors, with 95% significance bounds, as follows.

- ◇ Choose **Autocorrelations...** from the **Forecasting** submenu of **Analyze**. The **Autocorrelations** dialogue box will open.
- ◇ Enter **error** in the **Variables** field.
- ◇ In the **Display** area, make sure that **Autocorrelations** is checked, and click on **Partial autocorrelations** to deselect it. Unchecking **Partial Autocorrelations** will reduce the amount of unnecessary output.
- ◇ Leave all the check boxes in the **Transform** area unchecked.
- ◇ Click on **Options...** to open the **Autocorrelations: Options** dialogue box.
- ◇ Change the value in the **Maximum Number of Lags** field to 20, so that the sample autocorrelations for lags 1 to 20 will be displayed.
- ◇ In the **Standard Error Method** area, make sure that **Independence model** is selected. This determines the method that will be used to calculate the 95% significance bounds for the sample autocorrelations.
- ◇ Click on **Continue** to close the **Autocorrelations: Options** dialogue box.
- ◇ Click on **OK**.

A correlogram will be displayed at the end of the output in the **Viewer**, as shown in Figure 6.1.

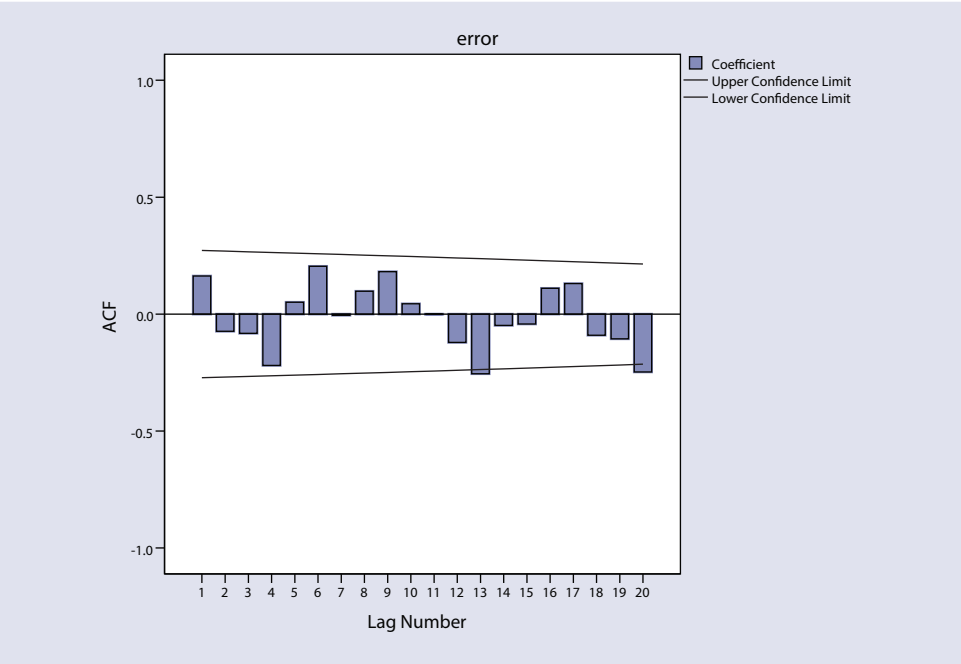


Figure 6.1 Correlogram for the forecast errors

Correlograms are called ACF plots in SPSS. (ACF stands for autocorrelation function.) Two lines are drawn on the correlogram; SPSS refers to them as ‘confidence limits’. In fact, they are not confidence limits but 95% *significance bounds*. If the sample autocorrelation at lag k , say, exceeds one of these bounds, this provides evidence against the null hypothesis that the underlying autocorrelation ρ_k is 0. Notice that the sample autocorrelations at lags 13 and 20 just exceed the significance bounds.

Notice also that the 95% significance bounds calculated by SPSS are not horizontal: they become closer together as the lag increases. In *Book 2*, 95% significance bounds were calculated using the formula $\pm 1.96/\sqrt{n}$, where n is the number of data values; this produces horizontal lines. However, SPSS calculates the bounds using a different method. When n is large, the two methods give similar results.

Immediately above the correlogram in the **Viewer** window, you will find the **Autocorrelations** output table shown on the next page.

The first column contains the lags (1 to 20). The second column contains the sample autocorrelations at lags 1 to 20. Look down this column. Notice that the autocorrelation at lag 13 is the largest in absolute value: $r_{13} = -0.255$; this is followed closely by the autocorrelation at lag 20: $r_{20} = -0.247$. The third column gives the standard errors of the autocorrelations: you may ignore this column.

The fourth column gives the values of the Ljung–Box test statistic. Notice that SPSS calls it the Box–Ljung test statistic. The p value for the test is given in the final column.

Recall that the Ljung–Box test is a test of the null hypothesis $\rho_1 = \rho_2 = \dots = \rho_k = 0$, so a value of the test statistic can be calculated for each value of k . SPSS gives all of the values up to the maximum lag that you specified (in this case, 20). You will need only the value of the test statistic corresponding to the maximum lag displayed in the correlogram, usually 20. In this case, the value of the test statistic is 25.576, with p value 0.180. This provides little evidence against the null hypothesis that the autocorrelations at lags 1 to 20 are zero.

Autocorrelations					
Series: error					
Lag	Autocorrelation	Std. Error ^a	Box-Ljung Statistic		
			Value	df	Sig. ^b
1	.163	.136	1.432	1	.231
2	-.073	.135	1.730	2	.421
3	-.082	.133	2.111	3	.550
4	-.219	.132	4.877	4	.300
5	.051	.130	5.029	5	.412
6	.205	.129	7.542	6	.274
7	-.005	.128	7.543	7	.375
8	.098	.126	8.144	8	.420
9	.181	.125	10.258	9	.330
10	.044	.123	10.386	10	.407
11	-.001	.122	10.386	11	.496
12	-.121	.120	11.403	12	.495
13	-.255	.119	16.029	13	.248
14	-.048	.117	16.199	14	.301
15	-.042	.115	16.334	15	.360
16	.110	.114	17.272	16	.368
17	.131	.112	18.627	17	.350
18	-.090	.110	19.293	18	.374
19	-.106	.109	20.237	19	.380
20	-.247	.107	25.576	20	.180

a. The underlying process assumed is independence (white noise).

b. Based on the asymptotic chi-square approximation.

Interpreting correlograms is *not* an easy task. The more you look at plots such as the one in Figure 6.1, the more you tend to imagine you can see in them. It is all too easy to over-interpret a correlogram, as the human brain is very good at picking out patterns. The Ljung–Box test provides a useful guide in this respect. Bearing this in mind, it is reasonable to conclude that this correlogram provides little evidence of departure from the white noise model.

Computer Activity 6.2 will give you some practice at obtaining and interpreting correlograms.

Computer Activity 6.2 Forecast errors for chemical temperatures

In Computer Activities 4.1 and 4.3, you used simple exponential smoothing to obtain forecasts for the temperature of a chemical process. The file **chemical2.sav** contains the output from the analysis in Computer Activity 4.3. Open the file now. There are three variables: **temperature**, **forecast** and **error**. The variable **error** contains the 1-step ahead forecast errors for the optimal value of the smoothing parameter α .

- (a) Obtain the correlogram for **error** for lags 1 to 20.
- (b) At what lags, if any, do the sample autocorrelations exceed the 95% significance bounds? Give the values of these sample autocorrelations, and interpret them.
- (c) Obtain the value of the Ljung–Box test statistic and the p value for lags 1 to 20. What do you conclude?
- (d) Summarize your findings. In your view, can the simple exponential smoothing method be improved upon?

Do not close the file **chemical2.sav**, as you will need it for the next activity.

In Computer Activity 6.3, a simple approach for investigating the assumption that the forecast errors are normally distributed is described.

Computer Activity 6.3 Histograms and normal curves

If the file **chemical2.sav** is not still open, then open it now. In Computer Activity 6.2, you investigated the autocorrelations between the forecast errors. The calculation of prediction intervals depends on the assumption that the forecast errors are normally distributed with mean zero and constant variance. One way of checking that the normality assumption is plausible is to obtain a histogram with a normal curve superimposed on it. This is done in SPSS using **Histogram...** from the **Legacy Dialogs** submenu of **Graphs**. SPSS calculates the sample mean and the sample variance, and uses these as estimates of the parameters of the normal distribution. Investigate the normality of the forecast errors, as follows.

- ◇ Choose **Histogram...** from the **Legacy Dialogs** submenu of **Graphs**. The **Histogram** dialogue box will open.
- ◇ Enter **error** in the **Variable** field.
- ◇ Select **Display normal curve**.
- ◇ Click on **OK**.

A histogram, with normal curve superimposed, will be displayed in the **Viewer** window, as shown in Figure 6.2.

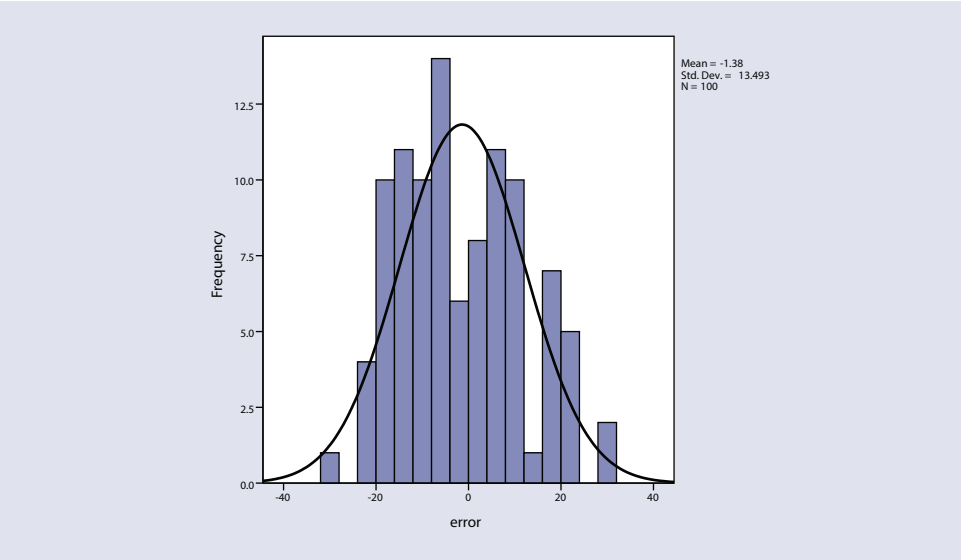


Figure 6.2 Histogram and normal curve

Note that the normal curve has been scaled so that the area under the curve is equal to the total number of observations. You can use this figure to assess whether the histogram plausibly represents observations from a normal distribution by looking at how well the histogram fits under the normal curve. Clearly, you would not expect anything like a perfect fit, owing to random variation. In this case, the correspondence between the curve and the histogram is not particularly good, but it does not provide strong evidence that a normal model is not adequate. So the assumption that the forecast errors are normally distributed is plausible.

Up to this point, the various steps in forecasting have been described separately. But it is important to conceive of forecasting as a process — that is, as an example of statistical modelling — in which a range of techniques and skills play their part. Most of the techniques discussed in Part II of *Book 2* can be implemented in SPSS. One difference is that SPSS automatically selects the starting values and the optimal smoothing parameters — so the methods for obtaining starting values described in *Book 2* are not required. A further

difference with *Book 2* is that SPSS does not calculate the SSE , though it can easily be obtained from the $RMSE$, which SPSS does give.

SPSS also calculates 95% prediction intervals for forecasted values — these are displayed below the forecasts in the **Forecast** output tables, **UCL** denoting the ‘upper confidence limit’ and **LCL** the ‘lower confidence limit’ (note however that these are not confidence limits, but prediction limits, as they relate to future observations, not parameter values). The method used by SPSS for calculating 1-step ahead prediction limits is slightly different from that used in *Book 2*, though the two methods should yield similar results when the number of observations is large.

Computer Activity 6.4 will give you some experience of the forecasting process, including the calculation of 1-step ahead prediction limits.

Computer Activity 6.4 Hong Kong vessel arrivals

The SPSS data file **tonnage.sav** contains the monthly time series of the total tonnage of all vessels (passenger and cargo) arriving at Hong Kong, from January 1997 to December 2003. Open the file now. There are four variables: **tonnage**, **YEAR_**, **MONTH_** and **DATE_**. The variable **tonnage** gives the net registered tonnage, in thousands of tonnes.

These data were obtained from the website of the Hong Kong Marine Department
<http://www.mardep.gov.hk/en/home.html> in April 2005.

- Obtain a time plot of **tonnage** by **DATE_**, and edit the plot so that the time axis labels appear only for January of each year. Describe the main features of the plot. In your view, does the time series display seasonal variation? Is an additive model appropriate for the time series?
- Obtain a forecast for the January 2004 tonnage (rounded to the nearest thousand tonnes) using the Holt–Winters exponential smoothing method.
- Write down the optimal parameter values and the corresponding $RMSE$. Calculate the SSE . Interpret the parameters in terms of the relative weights given to recent and past observations.
- Obtain a multiple time plot of the observed and forecasted values, and comment on how well the forecasts produced using this method agree with the observed values.
- Obtain and interpret the correlogram for the 1-step ahead forecast errors at lags 1 to 20. Carry out the Ljung–Box test. What do you conclude? Is the correlogram consistent with the forecast errors being white noise?
- Obtain a time plot and a histogram (with superimposed normal curve) of the 1-step ahead forecast errors. Use these to check the assumption that the forecast errors are normally distributed with mean zero and constant variance.
- Use the SSE you obtained in part (c) to calculate an approximate 95% prediction interval for the January 2004 tonnage. The actual value for January 2004 was 32 999. Comment on the accuracy of your forecast. Compare this prediction interval with that produced by SPSS.

Remember to tidy up the variables in **Variable View** before producing the plots for parts (e) and (f).

Summary of Chapter 6

In this chapter, you have learned how to obtain a correlogram with 95% significance bounds in SPSS, and how to carry out the Ljung–Box test. You have also learned how to obtain a histogram with a normal curve superimposed in order to check whether the forecast errors are normally distributed. You have applied these and other SPSS skills that you have acquired to undertake a complete analysis of a time series, thus illustrating the forecasting process.

Chapter 7

Stationary time series

In this chapter, you will learn how to transform time series to stationarity. Stationarity in variance can sometimes be produced by taking logarithms. The main tool for obtaining stationarity in mean is differencing. Stationarity in mean and in variance can be checked by inspecting a time plot. Time plots of time series produced by taking logarithms or by differencing (or both) can be obtained directly using **Sequence Charts...** from the **Forecasting** submenu of **Analyze**.

Computer Activity 7.1 Differencing time series

In Computer Activity 5.2, you applied Holt’s exponential smoothing method to the seasonally adjusted quarterly time series of the UK index of production. In this activity, you will learn how to use SPSS to difference this series and hence determine the order of differencing required to obtain a time series that is stationary in mean.

This time series and its first and second differences are discussed in Example 11.5 of *Book 2*.

The data are in the file **production.sav**. Open the file now. First, obtain a time plot of **index** by **DATE_**, such as that shown in Figure 7.1(a).

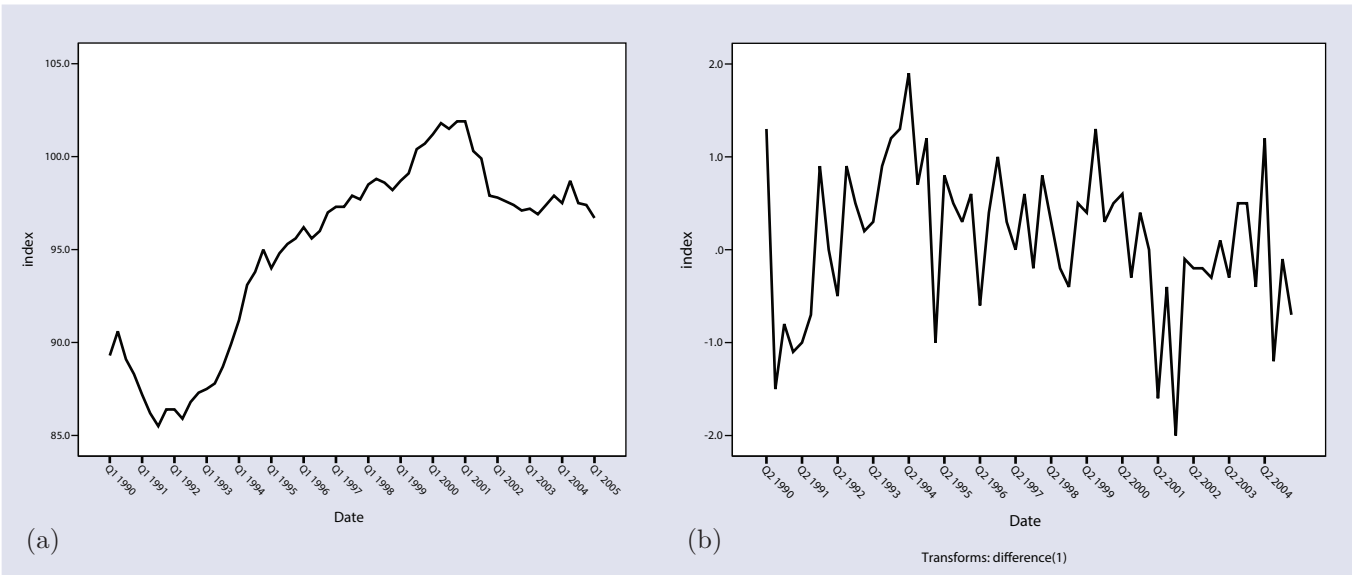


Figure 7.1 Time plots: (a) UK index of production (b) first differences

The time plot in Figure 7.1(a) shows that the level of the time series is not constant, so the time series is not stationary in mean.

Now obtain the time plot of first differences, as follows.

- ◇ Obtain the **Sequence Charts** dialogue box. Notice that **index** is still in the **Variables** field, and **DATE_** is in the **Time Axis Labels** field.
- ◇ In the **Transform** area, select **Difference** and check that 1 is in its field. Leave the **Natural log transform** and **Seasonally difference** check boxes unchecked.
- ◇ Click on **OK**.

The time plot of first differences will be displayed in the **Viewer** window. Although the vertical axis is labelled ‘index’, the variable plotted is the first difference of **index**, as indicated below the plot — see Figure 7.1(b).

This time plot suggests that the time series of first differences may not be stationary in mean: the trend appears to increase initially, then decline gradually.

The time plot of differences of order 2, or any higher order, can be obtained using the **Sequence Charts** dialogue box by changing the entry in the **Difference** field from 1 to the order desired. Obtain time plots of the second differences and third differences, such as those shown in Figure 7.2.

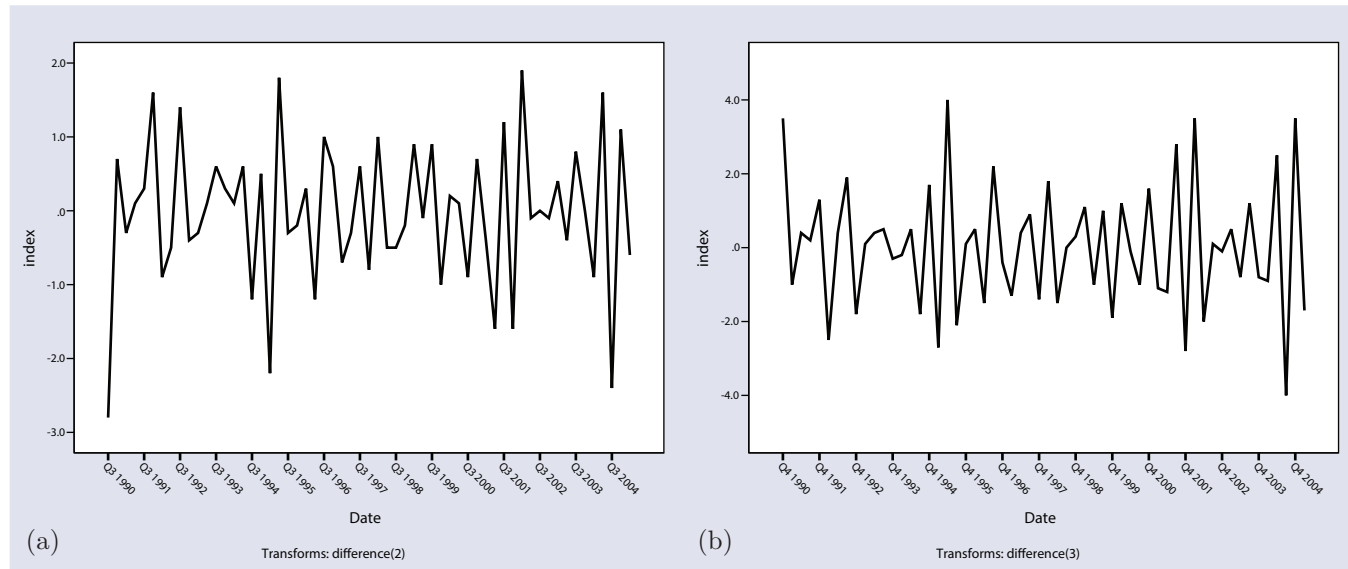


Figure 7.2 Time plots: (a) second differences (b) third differences

The time plot in Figure 7.2(a) shows that stationarity in mean is achieved by differencing twice: further differencing is unnecessary in this case. Since the size of the fluctuations is roughly constant, the time series of second differences is also stationary in variance.

Computer Activity 7.2 Share prices

The SPSS data file **ftse100.sav** contains the values of the FTSE100 share index at close of trade on the last day of each month between January 1988 and January 2005. The seasonality is small and may be ignored. Open the file now. The variable **ftse** contains the value of the monthly FTSE100 index, and **logftse** contains its logarithm.

- Obtain time plots of the first differences of the variables **ftse** and **logftse** by **DATE_**.
- For each of the time plots you obtained in part (a), decide whether the series shown is stationary in mean and in variance. Justify your decisions.

In Computer Activity 7.2, you found that the first differences of the variable **ftse** were not stationary in variance, but that the first differences of the logarithm of **ftse** were stationary in variance. Using a logarithmic transformation will not always produce stationarity in variance: sometimes other transformations are needed. However, checking whether a log transformation helps to produce stationarity in variance can be done directly in SPSS using the **Sequence Charts** dialogue box. This is illustrated in Computer Activity 7.3.

Computer Activity 7.3 Airline passengers

The SPSS data file **airline5.sav** contains data on the annual total number of airline passengers in the UK for each year between 1949 and 1999. Open the file now.

These data were used in Computer Activity 5.1 to obtain forecasts using Holt's method.

Obtain a time plot of **passengers** by **DATE_**, such as that shown in Figure 7.3(a).

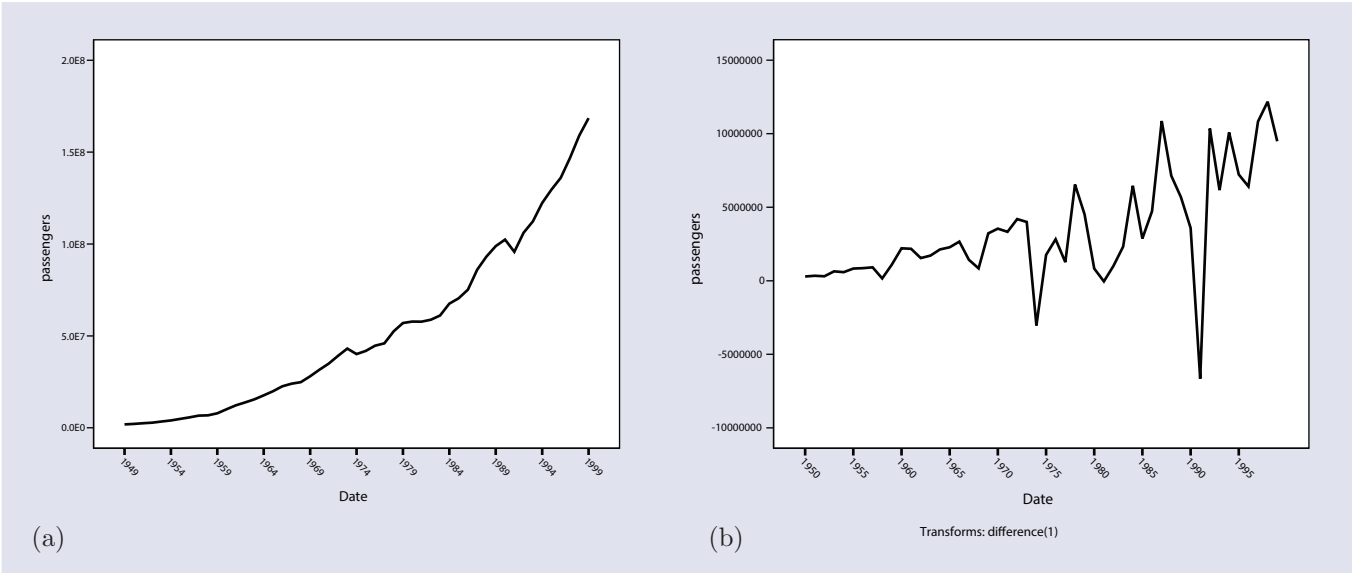


Figure 7.3 Airline passengers: (a) annual numbers (b) first differences

The time plot shows a markedly curved upward trend, so the time series is not stationary in mean. The scale of the data obscures the irregular component, so it is not possible to tell whether the time series is stationary in variance. However, if it is stationary in variance, then the time series of first differences will also be stationary in variance.

Obtain a time plot of first differences, such as that shown in Figure 7.3(b).

There is still a marked upward trend, and the fluctuations tend to increase in size as the level increases, so the time series of first differences is clearly not stationary either in mean or in variance. Thus the original series cannot be stationary in variance. Since differencing cannot produce a time series that is stationary in variance from a series that is not stationary in variance, a transformation is required. Once the series has been transformed to obtain stationarity in variance, stationarity in mean can be obtained by differencing (more than once, if necessary).

Transforming by taking natural logarithms and differencing can be done in a single operation when using the **Sequence Charts** dialogue box, so before trying other transformations it is worth checking whether the logarithmic transformation produces stationarity in variance. Do this now, as follows.

- ◇ Obtain the **Sequence Charts** dialogue box.

The previous settings will have been retained: **passengers** will be in the **Variables** field, and **DATE_** will be in the **Time Axis Labels** field; **Difference** will be checked, with 1 in its field.

- ◇ In the **Transform** area, check the **Natural log transform** box.
- ◇ Click on **OK**.

SPSS will calculate the logarithms of the numbers of passengers, then calculate the first differences of the logarithms, and display the time plot of the resulting time series in the **Viewer** window. A time plot of the first differences of the logarithms of the numbers of passengers is shown in Figure 7.4(a).

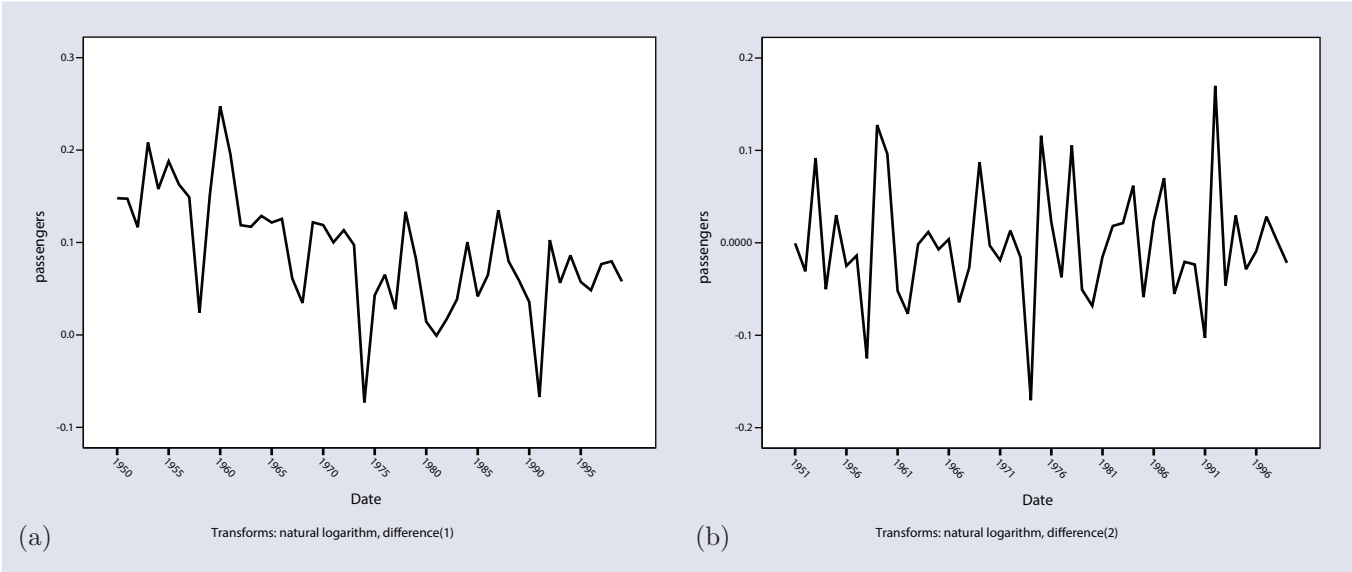


Figure 7.4 Logarithms of passenger numbers: (a) first differences (b) second differences

In this time plot, the size of the fluctuations is roughly constant, so the time series of first differences of the logarithms is stationary in variance. However, it is not stationary in mean, owing to the declining trend. Obtain a time plot of the series of second differences of the logarithms, as follows.

- Obtain the **Sequence Charts** dialogue box.

The previous settings will have been retained.

- In the **Transform** area, change the entry in the **Difference** field to 2.
- Click on **OK**.

A time plot of the second differences of logarithms will be displayed in the **Viewer** window. A time plot is shown in Figure 7.4(b); this shows that the time series of second differences of the logarithms of passenger numbers is stationary in both mean and variance.

Computer Activity 7.4 will give you some practice at transforming a time series to obtain a time series that is stationary in both mean and variance.

Computer Activity 7.4 Electricity demand

The SPSS data file **electricity.sav** contains data on the monthly electricity demand at a utility in California, between January 1965 and December 1989. Open the file now. The variable **sademand** contains the seasonally adjusted demand, in megawatt-hours.

- Obtain a time plot of **sademand** by **DATE_**. Describe its two main features.
- Transform this time series to stationarity, using the log transformation and differencing.

These data were described in Computer Activity 5.4.

You may assume that **sademand** is non-seasonal.

Summary of Chapter 7

In this chapter, you have learned how to difference a time series in SPSS. In some cases, it is also necessary to transform a time series to obtain stationarity in variance. You have learned how to combine the log transformation and differencing in SPSS to obtain a time series that is stationary in both mean and variance.

Chapter 8

ARIMA modelling in SPSS

The steps involved in a forecasting exercise using ARIMA modelling are discussed in Section 14 of *Book 2*. In this chapter, you will learn how to use SPSS to carry out these steps. Selecting an ARIMA model (after using a log transformation if required) is discussed in Section 8.1. Fitting an ARIMA model and checking its adequacy are described in Section 8.2. In Section 8.3, you will learn how to obtain forecasts based on an ARIMA model, and prediction intervals for the forecasts.

8.1 Selecting an ARIMA model

The first step in selecting an ARIMA model for a time series is to check that the time series is stationary. If it is not stationary, then you will need to transform it to obtain a stationary series, by differencing and, if necessary, using other transformations such as the log transformation. Transforming time series to produce stationarity using the log transformation and differencing was described in Chapter 7. The minimum number of differences required to obtain a stationary series is the order of differencing, d , of the ARIMA model.

The next step in selecting an ARIMA model is to identify appropriate values for p and q , the orders of the autoregressive and moving average components of the model for the stationary series. This is done by inspecting the correlogram and the partial correlogram for the stationary series. Correlograms and partial correlograms are obtained using **Autocorrelations...** from the **Forecasting** submenu of **Analyze**. Selecting an ARIMA model is illustrated in Computer Activity 8.1.

Computer Activity 8.1 Selecting a model for the FTSE100 index

The SPSS data file **ftse100.sav** contains the values of the FTSE100 index at close of trade on the last day of each month between January 1988 and January 2005. Open the file now.

In Computer Activity 7.2, you found that applying the log transformation and differencing once produces a time series that is stationary in both mean and variance. Obtain the correlogram and the partial correlogram for the stationary series, as follows.

- ◇ Choose **Autocorrelations...** from the **Forecasting** submenu of **Analyze**. The **Autocorrelations** dialogue box will open.
- ◇ Enter **ftse** in the **Variables** field.
- ◇ In the **Transform** area, check **Natural log transform** and **Difference**, and make sure that 1 appears in the **Difference** field. Leave the **Seasonally difference** box unchecked.

These choices ensure that the sample ACF and sample PACF will be calculated for the first differences of the logarithms of the time series in the variable **ftse**.

- ◇ In the **Display** area, make sure that both **Autocorrelations** and **Partial Autocorrelations** are checked.
- ◇ Click on **Options** to open the **Autocorrelations: Options** dialogue box.
- ◇ Enter 20 in the **Maximum Number of Lags** field.
- ◇ Click on **Continue** to close the **Autocorrelations: Options** dialogue box.
- ◇ Click on **OK**.

Several output tables will be displayed in the **Viewer** window, as well as the correlogram and the partial correlogram for the stationary series. Scroll down the output until you come to the correlogram, which is shown in Figure 8.1.

Entering **logftse** in the **Variables** field and leaving the **Natural log transform** box unchecked would produce the same results.

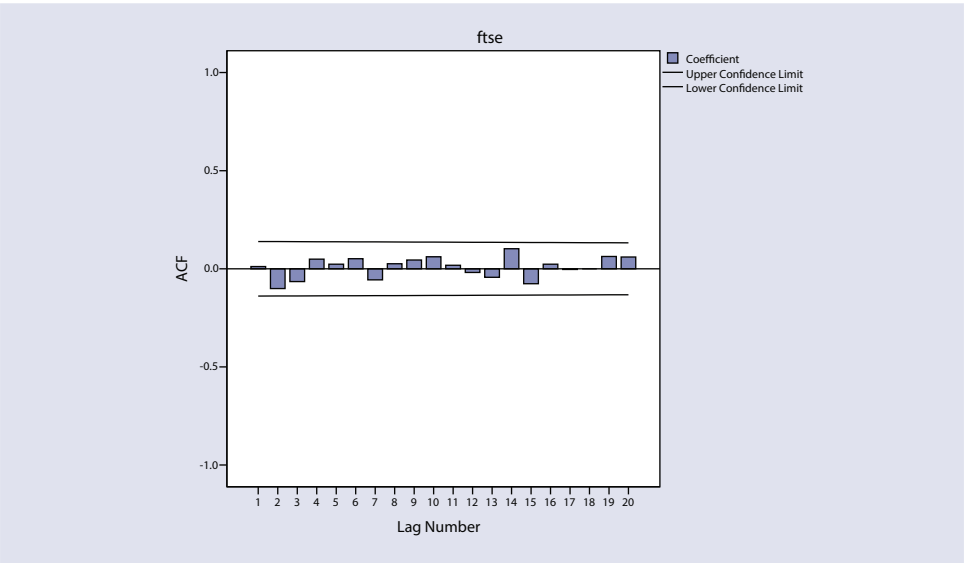


Figure 8.1 Correlogram

None of the sample autocorrelations at lags 1 to 20 exceeds the significance bounds. Thus it would appear that there is little evidence of any non-zero autocorrelation at these lags. You may check this further by looking at the value of the Ljung–Box test statistic and the p value for the test; these are given in the **Autocorrelations** output table immediately above the correlogram in the **Viewer** window. The value of the test statistic is 12.286, and the p value is 0.906. So there is little evidence against the null hypothesis of zero autocorrelation at lags 1 to 20.

The Ljung–Box test is discussed in Computer Activity 6.1.

Now scroll down the output until you reach the partial correlogram, which SPSS calls the Partial ACF. This is shown in Figure 8.2.

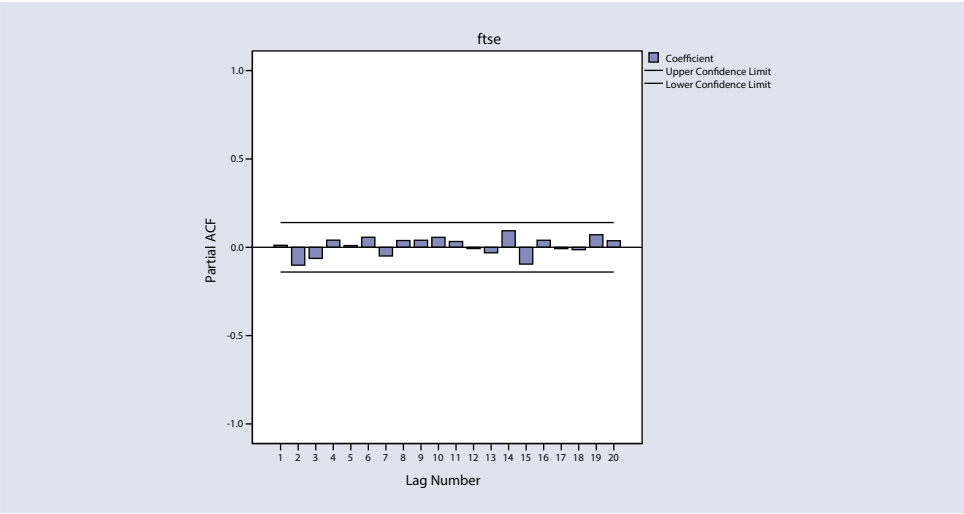


Figure 8.2 Partial correlogram

The horizontal lines on either side of the central line are the 95% significance bounds; these are drawn at $\pm 1.96/\sqrt{n}$, where n is the length of the time series. As for the significance bounds on the correlogram, these are incorrectly described as confidence limits.

See Computer Activity 6.1.

The partial correlogram suggests that there are no non-zero partial autocorrelations at lags 1 to 20. Note that no significance test (like the Ljung–Box test for zero autocorrelations) is available in SPSS to test the null hypothesis of zero partial autocorrelations.

The conclusion is that the stationary series is uncorrelated, so that $p = q = 0$. Hence, since the order of differencing is 1, a suitable model for the logarithm of `ftse` is an ARIMA(0, 1, 0) model.

See Table 14.1 of *Book 2*.

Computer Activity 8.2 *Selecting a model for the UK index of production*

The SPSS data file `production.sav` contains values of the seasonally adjusted quarterly time series of the UK index of production in the variable `index`. Open the file now. In Computer Activity 7.1, it was suggested that two differencing steps are required to produce a time series that is stationary in both mean and variance. Thus d , the order of differencing required to obtain a stationary series, is 2.

Use the method described in Computer Activity 8.1 to obtain the correlogram and the partial correlogram for lags 1 to 20 for the differenced series. In the **Transform** area of the **Autocorrelations** dialogue box, you should check the **Difference** box and enter 2 in its field, but leave the other boxes unchecked.

Analyze > Forecasting > Autocorrelations...

With these choices, the correlogram and the partial correlogram for the second differences of `index` will be obtained. These are shown in Figure 8.3.

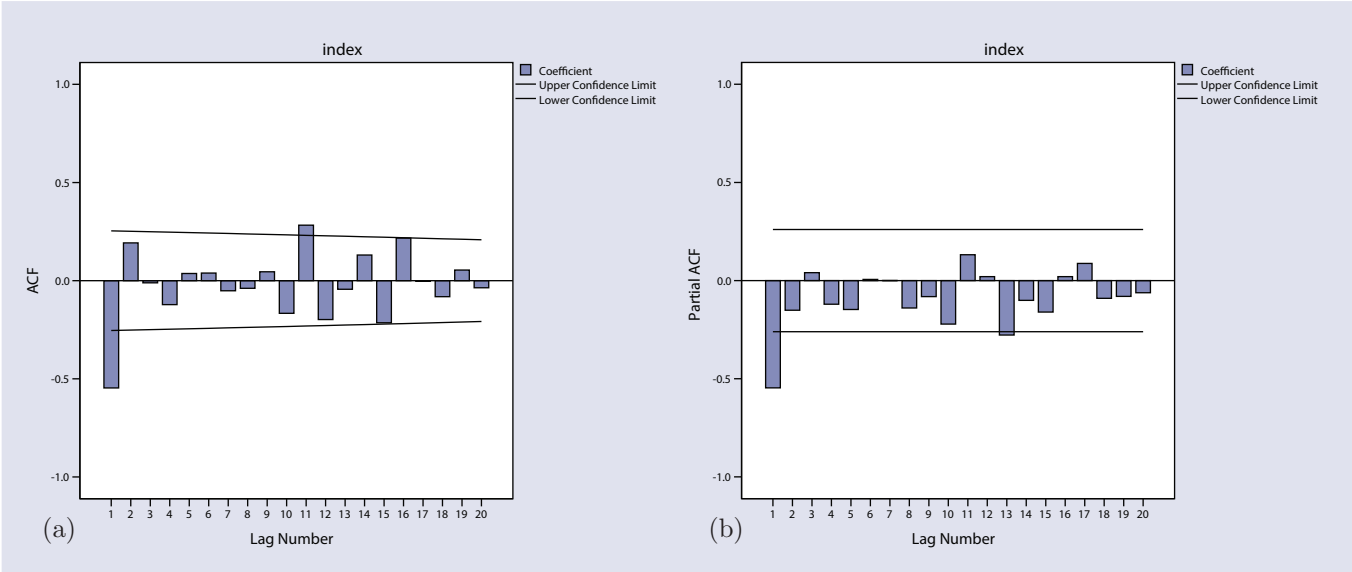


Figure 8.3 Second differences: (a) correlogram (b) partial correlogram

Note that the significance bounds on the correlogram get slightly smaller (in absolute value) with increasing lag, whereas those on the partial correlogram do not. This difference occurs because SPSS uses different methods for calculating the two sets of bounds. This difference is not important (and is visible only for relatively short time series).

Both the sample ACF and the sample PACF show a large negative value at lag 1, and much smaller values at higher lags. This suggests two plausible models for the differenced series: an AR(1) model and a MA(1) model. Thus, since the order of differencing is 2, two plausible models for the original time series are an ARIMA(1, 2, 0) model and an ARIMA(0, 2, 1) model. Note that the principle of parsimony is of no help here, since $p + q = 1$ for both models.

The principle of parsimony is discussed in Subsection 14.2 of *Book 2*.

Computer Activity 8.3 will give you some practice at identifying an ARIMA model.

Computer Activity 8.3 *Selecting a model for the airline passenger data*

The SPSS data file **airline5.sav** contains the time series of annual numbers of airline passengers in the UK between 1949 and 1999. Open the file now. In Computer Activity 7.3, you found that a log transformation and differencing twice produces a time series that is stationary in both mean and variance.

- Obtain the correlogram and the partial correlogram for lags 1 to 20 for the second differences of the time series of logarithms of the annual numbers of airline passengers. The annual numbers are in the variable **passengers**.
- Suggest one or more suitable models for the stationary series, based on the appearance of the sample ACF and the sample PACF. (You will find Table 14.1 of *Book 2* useful in deciding which models may be suitable.)
- Using the principle of parsimony, if appropriate, identify a single suitable ARIMA model for the logarithm of the number of passengers.

8.2 Fitting and checking ARIMA models

In this section, you will learn how to fit an ARIMA model in SPSS, and how to check that the model is adequate. Fitting a model is done using **Create Models...** from the **Forecasting** submenu of **Analyze**. This includes a facility for using a log transformation before fitting an ARIMA model. Checking the fit of a model is done by producing appropriate graphical displays. Fitting and checking an ARIMA model to the UK index of production is illustrated in Computer Activities 8.4 and 8.5.

Computer Activity 8.4 *Fitting an ARIMA model*

Values of the seasonally adjusted quarterly UK index of production are in the SPSS data file **production.sav**. Open the file now. In Computer Activity 8.2, it was suggested that an ARIMA(0, 2, 1) model might be suitable for this time series. Fit an ARIMA(0, 2, 1) model to the time series, as follows.

The first step is to define the model, as follows.

- ◇ Choose **Create Models...** from the **Forecasting** submenu of **Analyze**. The **Time Series Modeler** dialogue box will open.
- ◇ Enter **index** in the **Dependent Variables** field in the **Variables** panel. Leave the **Independent Variables** field empty.
- ◇ Select **ARIMA** from the **Method** drop-down list.
- ◇ Click on the **Criteria...** button. The **Time Series Modeler: ARIMA Criteria** dialogue box will open. Make sure the **Model** panel is uppermost. (You will not use the **Outliers** panel.)

Now define the ARIMA model in the **ARIMA Orders** area by entering the values $p = 0$, $d = 2$ and $q = 1$ under **Structure**, as follows.

- ◇ Leave 0 in the **Autoregressive (p)** field in the **Nonseasonal** column.
- ◇ Enter 2 in the **Difference (d)** field in the **Nonseasonal** column.
- ◇ Enter 1 in the **Moving Average (q)** field in the **Nonseasonal** column.
- ◇ The fields in the **Seasonal** column relate to seasonal ARIMA models, so leave them empty.

Seasonal ARIMA models are not covered in M249.

The model definition is completed as follows.

- ◇ No transformation is required, so leave **None** selected in the **Transformation** area.
- ◇ Make sure that **Include constant in model** is checked. When this is checked, a constant μ , representing the mean of the differenced series, is fitted. In M249, a constant will always be fitted.
- ◇ Click on **Continue**.

This completes the definition of the ARIMA model. The rest of the set-up involves defining the output required and is similar to that described for exponential smoothing. In this activity, values of the time series beyond the range of the data will not be forecasted.

- ◇ Click on the **Statistics** tab.
- ◇ Make sure that the **Display fit measures, Ljung–Box statistic, and number of outliers by model** box is checked. Select **Root mean square error** and **Parameter estimates**, and deselect any other options that are selected. Since future values will not be forecasted, leave the **Display forecasts** box unchecked.

The forecasts and the 1-step ahead prediction errors should be added to the data file, as follows.

- ◇ Click on the **Save** tab.
- ◇ In the **Save Variables** area, select **Predicted Values** and **Noise Residuals** in the **Save** column.
- ◇ Click on the **Plots** tab.
- ◇ Deselect any options that are checked. Plots will be discussed in the next activity, in the context of model checking.

Since future values will not be forecasted in this activity, this completes the set-up.

- ◇ Click on **OK** to run the model.

The output in the **Viewer** window includes three output tables. The first is labelled **Model Description**. Note that the model is described as ARIMA(0, 2, 1)(0, 0, 0). The last bracket (0, 0, 0) refers to the seasonal component of the model, of which there is none. You can ignore this bracket: the model will be referred to simply as ARIMA(0, 2, 1).

The next output table is labelled **Model Statistics**. The only item to consider here is the value of the *RMSE*, which is 0.826. You should ignore the other entries in this table. In particular, the Ljung–Box test of zero autocorrelation will be carried out separately, and always on 20 autocorrelations (here it is performed on 18).

The relationship between the *RMSE* and the *SSE* is discussed at the end of the activity.

The third output table is the **ARIMA Model Parameters** output table, reproduced below.

ARIMA Model Parameters				Estimate	SE	t	Sig.
index-Model_1	index	No Transformation	Constant	-.001	.028	-.019	.985
			Difference	2			
		MA	Lag 1	.746	.094	7.909	.000

The parameter estimates are given in this table. Recall that the model fitted to the twice-differenced series Y_t is

$$Y_t - \mu = Z_t - \theta_1 Z_{t-1}.$$

The MA(1) model is defined in Subsection 13.1 of *Book 2*.

The first row of the table relates to the estimated value of the constant, which is $\hat{\mu} = -0.001$. The standard error of the estimates is in the third column, labelled **SE**. So, for example, the standard error of $\hat{\mu}$ is 0.028. You may ignore the last two columns of the table, which relate to a significance test of the null hypothesis that $\mu = 0$.

The second row of the table gives the order of differencing, $d = 2$. This value was specified, so it is not really an estimate at all.

The estimate of θ_1 is in the third row of the table, labelled **MA Lag 1**. The estimate is $\hat{\theta}_1 = 0.746$ and the standard error of $\hat{\theta}_1$ is 0.094. Again, you may ignore the last two columns of the table, which relate to a significance test of the null hypothesis that $\theta_1 = 0$.

Thus, the model for the second differences of the index of production is

$$Y_t + 0.001 = Z_t - 0.746Z_{t-1}.$$

Finally, look at the variables in the **Data Editor**. Two new variables have been created. The variable **Predicted_index_Model_1** contains the 1-step ahead forecasts for the index of production: these are the values predicted by the estimated model, and ‘integrated’ (that is, summed) so that they relate to the original series, not the differenced data. The variable **NResidual_index_Model_1** contains the 1-step ahead forecast errors. Note that the 1-step ahead forecast errors are equal to the data values (in **index**) minus the 1-step ahead forecasts. The first two values of both new variables are missing (as indicated by the dots in rows 1 and 2 in the **Data View** panel) because the order of differencing is 2.

For this reason, for ARIMA models, the relationship between the *RMSE* and the *SSE* is based on the number of observations after differencing, which is $n - d$. Thus:

$$RMSE = \sqrt{\frac{SSE}{n - d - k}}, \quad SSE = (n - d - k) \times RMSE^2,$$

where n is the number of observations, d is the order of differencing, and k is the number of estimated parameters. For the index of production data, $n = 61$, $d = 2$ and $k = 2$ (there are two estimated parameters, μ and θ_1). So the *SSE* is

$$SSE = (61 - 2 - 2) \times 0.826^2 \simeq 38.9.$$

For comparing different ARIMA models, either the *RMSE* or the *SSE* can be used. When the models to be compared have different numbers of parameters, it is preferable to use the *RMSE*.

Having fitted an ARIMA model, the next step is to check whether the model is adequate, as follows. First, check that the 1-step ahead forecasts are close to the original time series. Then, check that the forecast errors are normally distributed with mean zero and constant variance. And finally, check that the forecast errors are uncorrelated. This is illustrated in Computer Activity 8.5 for the model fitted in Computer Activity 8.4.

Computer Activity 8.5

Checking the model

The results of Computer Activity 8.4 have been saved in the SPSS data file **production2.sav**. Open this file and edit it to avoid clutter, as follows. Change the variable name **Predicted_index_Model_1** to **forecast** and the variable name **NResidual_index_Model_1** to **error**, and delete the labels for these two variables.

Use **Sequence Charts...** from the **Forecasting** submenu of **Analyze** to obtain a multiple time plot of the data and the 1-step ahead forecasts, such as that shown in Figure 8.4 (overleaf).

A similar multiple time series plot could also have been obtained in the previous activity using the **Plots** panel of the **Time Series Modeler** dialogue box.

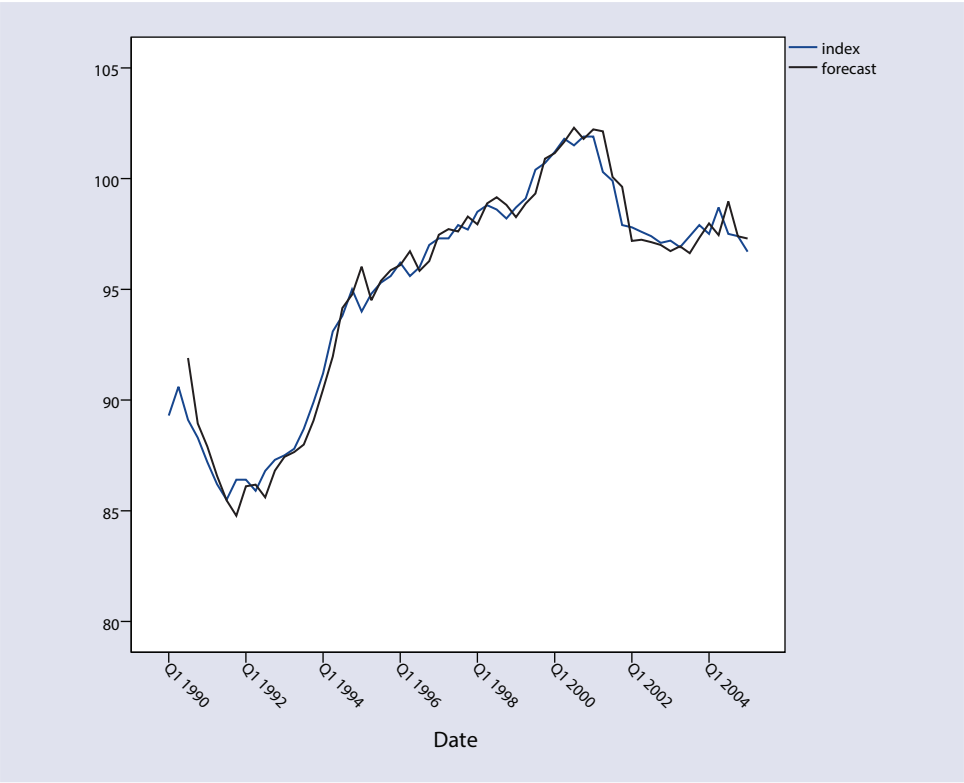


Figure 8.4 Observed and forecasted values of UK index of production

This time plot shows that the forecasts (the fitted values) match the observed values quite closely.

Next check that the distribution of the forecast errors (in the variable `error`) is approximately normal with mean zero and constant variance: obtain a time plot of the errors and a histogram of the errors with a superimposed normal curve, such as those shown in Figure 8.5.

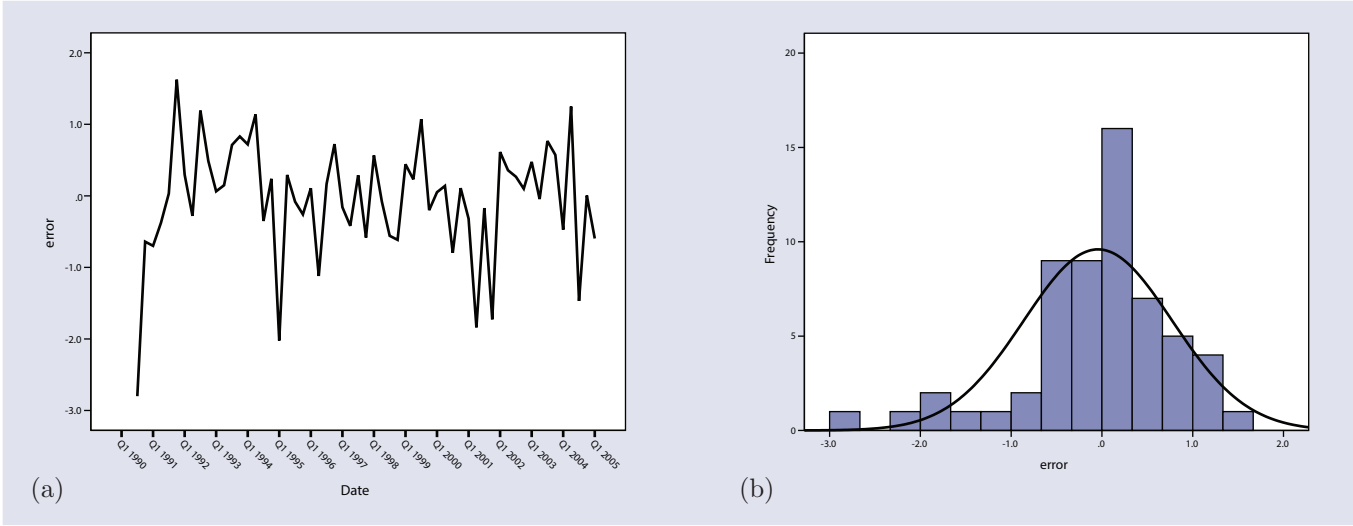


Figure 8.5 Forecast errors: (a) time plot (b) histogram

The time plot in Figure 8.5(a) shows that the forecast errors fluctuate around zero with roughly constant variance, as required.

The histogram in Figure 8.5(b) has a rather long tail on the left, which suggests that the forecast errors may be negatively skewed, and hence not normally distributed. However, the skewness is not marked enough for the normality assumption to be untenable.

Finally, check that the forecast errors are uncorrelated by obtaining the correlogram for the forecast errors for lags 1 to 20. The correlogram is shown in Figure 8.6.

Analyze > Forecasting > Autocorrelations...

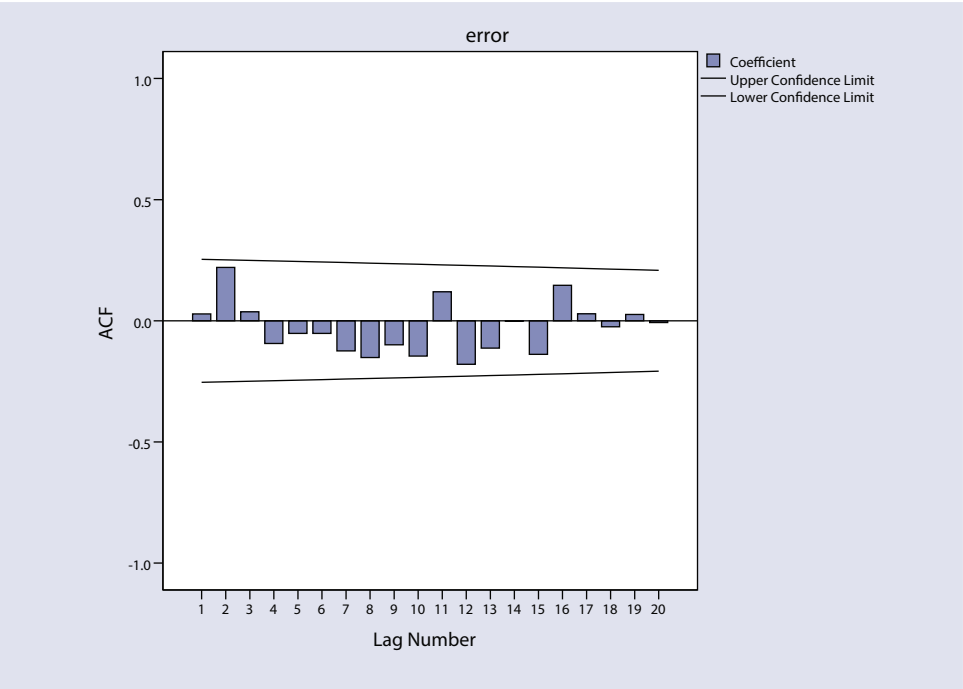


Figure 8.6 Correlogram for forecast errors

None of the sample autocorrelations exceeds the significance bounds, and there is no obvious pattern to them. Thus there is no reason to suspect that any of the autocorrelations are non-zero at lags 1 to 20.

The results of the Ljung–Box test of zero autocorrelation at lags 1 to 20 are given in the Autocorrelations output table above the correlogram. The value of the test statistic is 17.192, with p value 0.640. Thus there is little evidence against the null hypothesis of zero autocorrelation at lags 1 to 20.

Overall, the ARIMA(0, 2, 1) model appears to be adequate, though the assumption that the errors are normally distributed may be questionable.

In Computer Activity 8.2, the ARIMA(0, 2, 1) model and the ARIMA(1, 2, 0) model were suggested as possible models for the UK index of production data. In Computer Activity 8.6, you will investigate the second of these models.

Computer Activity 8.6 Fitting another ARIMA model

Open the file **production.sav**.

- (a) Fit an ARIMA(1, 2, 0) model to the data. Obtain the *RMSE* and the estimates of the parameters of the model: β_1 and μ . (The parameter β_1 is labelled AR1 in the **Parameter Estimates** table.) Calculate the *SSE*.
- (b) Write down the model formula for the time series of second differences.
- (c) Check the adequacy of the ARIMA(1, 2, 0) model.
- (d) Briefly compare the fit and adequacy of this ARIMA(1, 2, 0) model and the ARIMA(0, 2, 1) model that was fitted in Computer Activity 8.4 and assessed in Computer Activity 8.5.

An example of fitting an ARIMA model to a time series that has been transformed by taking logarithms is given in Computer Activity 8.7.

Computer Activity 8.7 An ARIMA model for the logarithm of electricity demand

The SPSS data file **electricity.sav** contains data on the monthly electricity demand at a utility in California, between January 1965 and December 1989. Open the file now. The variable **sademand** contains the seasonally adjusted demand, in megawatt-hours.

These data were described in Computer Activity 5.4.

In Computer Activity 7.4, you found that taking logarithms and differencing once results in a stationary time series. Further analysis suggests that an ARIMA(1, 1, 1) model fitted to the logarithm of **sademand** might be appropriate. In this activity, you will apply the log transformation and fit this model.

- ◇ Obtain the **Time Series Modeler** dialogue box.

Analyze > Forecasting >
Create Models...

First, define the ARIMA model for the log of **sademand**, as follows.

- ◇ In the **Variables** panel, enter **sademand** in the **Dependent Variables** field. Leave the **Independent Variables** field empty.
- ◇ Select **ARIMA** from the **Method** drop-down list.
- ◇ Click on the **Criteria...** button. The **Time Series Modeler: ARIMA Criteria** dialogue box will open.
- ◇ Under **Structure** in the **ARIMA Orders** area, enter 1 in the **Autoregressive (p)** field in the **Nonseasonal** column.
- ◇ Enter 1 in the **Difference (d)** field in the **Nonseasonal** column.
- ◇ Enter 1 in the **Moving Average (q)** field in the **Nonseasonal** column.
- ◇ The fields in the **Seasonal** column relate to seasonal ARIMA models, so leave them empty.
- ◇ In the **Transformation** area, select **Natural log**.
- ◇ Make sure that **Include constant in model** is checked.
- ◇ Click on **Continue**.

Seasonal ARIMA models are not covered in M249.

This completes the definition of the ARIMA model. The rest of the set-up is as before.

- ◇ Click on the **Statistics** tab.
- ◇ Make sure that the **Display fit measures**, **Ljung-Box statistic**, and **number of outliers by model** box is checked. Select **Root mean square error** and **Parameter estimates**, and deselect any other options that are selected. Since future values will not be forecasted, leave the **Display forecasts** box unchecked.
- ◇ Now click on the **Plots** tab.
- ◇ Leave unchecked all the boxes in the **Plots for Comparing Models** area. In the **Plots for Individual Models** area, select **Series**. In the **Each Plot Displays** sub-area, select **Observed values** and **Fit values**, and deselect **Forecasts**.

To add the forecasts and the 1-step ahead prediction errors to the data file, proceed as follows.

- ◇ Click on the **Save** tab.
- ◇ In the **Save Variables** area, select **Predicted Values** and **Noise Residuals** in the **Save** column.

Since future values will not be forecasted in this activity, this completes the set-up.

- ◇ Click on **OK** to run the model.

The *RMSE* and the model parameters are given in the **Model Statistics** output table and the **ARIMA Model Parameters** output table in the **Viewer** window.

The *RMSE* is 2799.860; note that this is calculated from the original data, not the log transformed data. The parameter estimates are $\hat{\beta}_1 = 0.363$, $\hat{\theta}_1 = 0.861$ and $\hat{\mu} = 0.005$. Hence the fitted model for the differenced series is

$$Y_t - 0.005 = 0.363(Y_{t-1} - 0.005) + Z_t - 0.861Z_{t-1}.$$

The adequacy of the model may be checked in exactly the same way as when a log transformation has not been used. Start by obtaining a multiple time plot of *sademand* and the 1-step ahead forecasts, such as that shown in Figure 8.7.

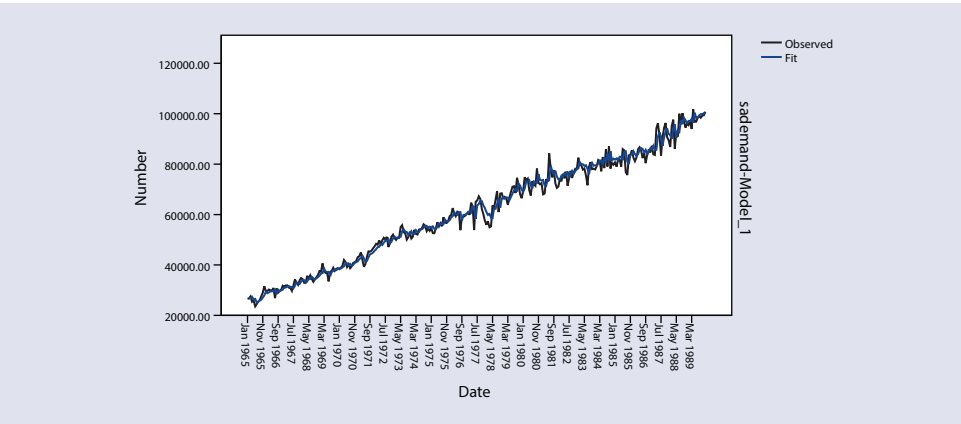


Figure 8.7 Observed and forecasted values

This does not reveal any problems with the fit of the model. The next step is to check the distribution of the forecast errors. Look in the **Data View** panel of the **Data Editor**. The newly created variable *Predicted_sademand_Model_1* contains the 1-step ahead forecasts, transformed so that they are expressed on the same scale as the variable *sademand*. In the **Variable View** panel, change the name of this variable to the more manageable *forecast* and delete its label. The second new variable *NResidual_sademand_Model_1* contains the 1-step ahead forecast errors of the logged and differenced series, which may be used to check the adequacy of the model. Change the name of this second new variable to *error* and delete its label.

Now check that the distribution of the forecast errors is approximately normal with mean zero and constant variance by obtaining a time plot of the forecast errors and a histogram of the errors with a normal curve superimposed.

A time plot and histogram are shown in Figure 8.8.

Use the **Sequence Charts** dialogue box with all the boxes in the **Transform** area left unchecked. Note that the multiple time plot of *sademand* and *forecast* can also be produced in this way.

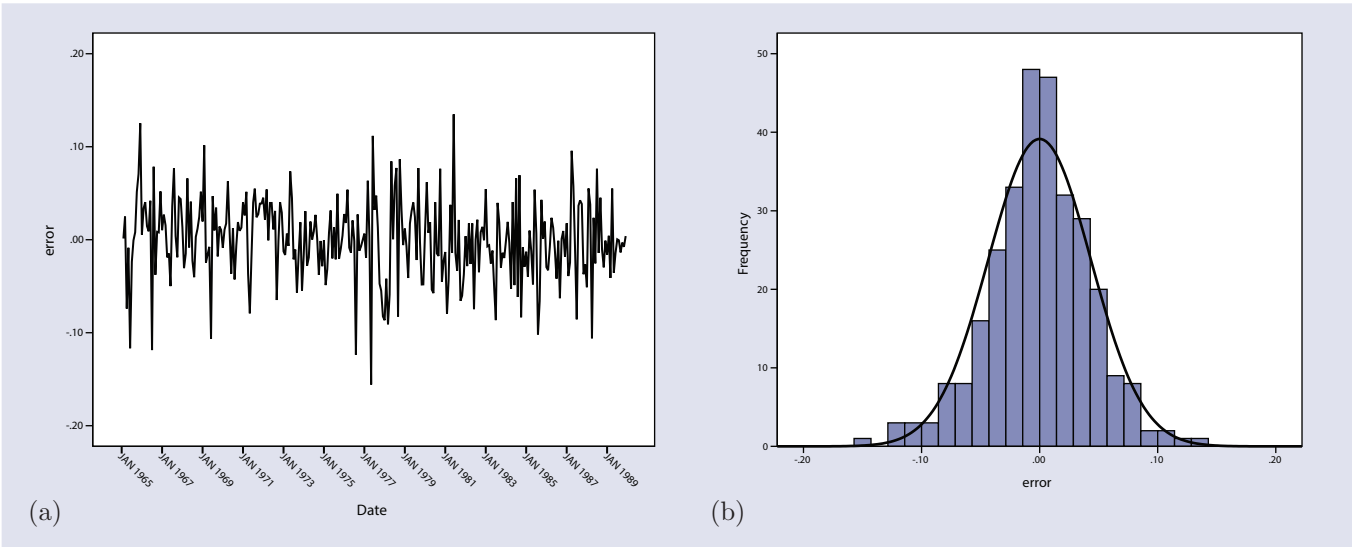


Figure 8.8 Forecast errors: (a) time plot (b) histogram

The time plot in Figure 8.8(a) suggests that the errors are distributed with mean zero and constant variance. The histogram in Figure 8.8(b) suggests that the forecast errors are approximately normally distributed, as required.

Finally, check that the forecast errors are uncorrelated by obtaining the correlogram for lags 1 to 20, as shown in Figure 8.9.

Use the **Autocorrelations** dialogue box with all the boxes in the **Transform** area unchecked.

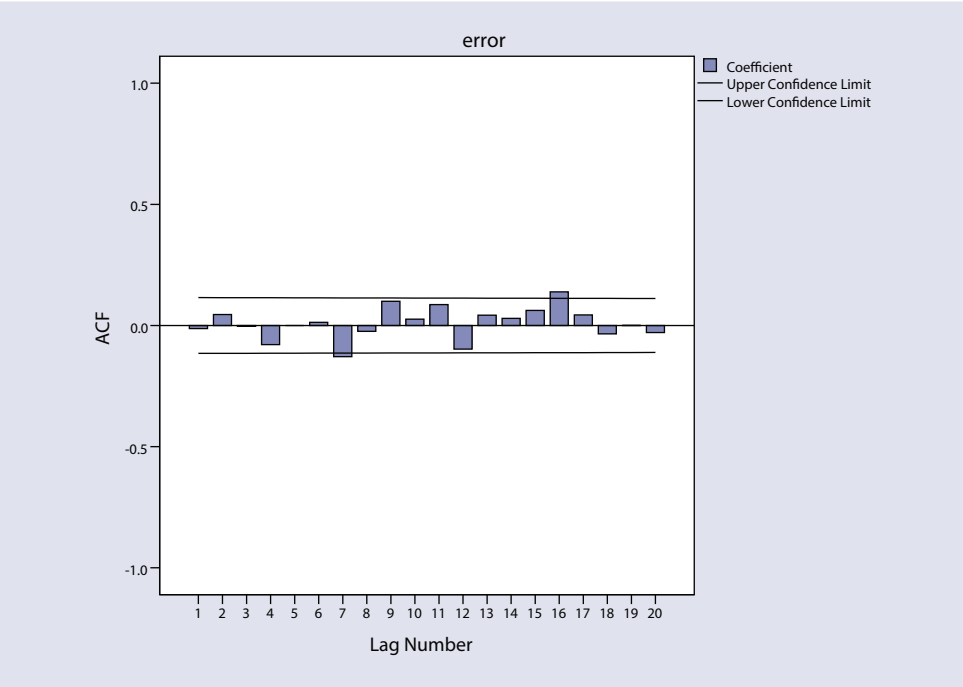


Figure 8.9 Correlogram for forecast errors

Two sample autocorrelations (at lags 7 and 16) just exceed the significance bounds. However, the value of the Ljung–Box test statistic is 26.021, with $p = 0.165$, so there is little evidence against the null hypothesis of zero autocorrelation at lags 1 to 20.

Thus the ARIMA(1, 1, 1) model provides an adequate fit to the time series of logarithms of the seasonally adjusted electricity demand.

Computer Activity 8.8 will give you some more practice at fitting ARIMA models.

Computer Activity 8.8 *An ARIMA model for airline passenger numbers*

The SPSS data file **airline5.sav** contains the time series of annual numbers of airline passengers in the UK between 1949 and 1999. Open the file now. In Computer Activity 7.3, it was found that a log transformation and differencing twice produces a stationary series. In Computer Activity 8.3, you found that, after using a log transformation, an ARIMA(0, 2, 1) model may be appropriate for these data.

- (a) Fit an ARIMA(0, 2, 1) model to the airline data. Obtain the *RMSE* and the estimated values of θ_1 and μ .
- (b) Obtain a multiple time plot of the observed and forecasted values, a time plot of the forecast errors, a histogram of the forecast errors with a normal curve superimposed, and the correlogram for the forecast errors for lags 1 to 20. Also obtain the value of the Ljung–Box test statistic and the p value for the test of zero autocorrelation of the forecast errors at lags 1 to 20.
- (c) Is the model adequate for these data? Briefly summarize your conclusions.

8.3 Forecasting with ARIMA models

In this section, you will learn how to use SPSS to obtain forecasts of future values of a time series, and prediction intervals for these forecasts. The method for obtaining prediction intervals for 1-step ahead forecasts obtained from an ARIMA model is similar to that described in Section 9 of *Book 2* for exponential smoothing methods. (The details will be omitted.) In fact, SPSS calculates predictions and prediction intervals for any number of steps ahead.

Forecasts and prediction intervals may be obtained when using **Create Models...** from the **Forecasting** submenu of **Analyze** to fit an ARIMA model. This is illustrated in Computer Activity 8.9.

Computer Activity 8.9 Forecasting the UK index of production

The SPSS data file **production.sav** contains the seasonally adjusted quarterly time series of the UK index of production. The time series spans the period from the first quarter of 1990 to the first quarter of 2005. Open the file now or make it active if it is already open. In Computer Activity 8.4, the ARIMA(0, 2, 1) model was fitted to the data. Obtain the forecasted values until the first quarter of 2010, and 95% prediction limits for the forecasted values, as follows.

- ◇ Obtain the **Time Series Modeler** dialogue box.
- ◇ Click on **Reset** to restore default settings.
- ◇ In the **Variables** panel, enter **index** in the **Dependent Variables** field. Leave the **Independent Variables** field empty.
- ◇ Use the **Method** drop-down list and the **Time Series Modeler: ARIMA Criteria** dialogue box to specify the ARIMA(0, 2, 1) model.

Analyze > Forecasting >
Create Models...

See Computer Activity 8.4.

Now continue with the set-up, taking care to specify the output plots required.

- ◇ Click on the **Statistics** tab.
- ◇ Make sure that the **Display fit measures, Ljung–Box statistic, and number of outliers by model** box is checked. Select **Root mean square error** and **Parameter estimates**, and deselect any other options that are selected.
- ◇ Leave the **Display forecasts** box unchecked, as the **Forecast** output table would be too wide.
- ◇ Now click on the **Plots** tab.
- ◇ Leave unchecked all the boxes in the **Plots for Comparing Models** area. In the **Plots for Individual Models** area, select **Series**.
- ◇ In the **Each Plot Displays** sub-area, select all five options: **Observed values, Forecasts, Fit values, Confidence intervals for forecasts** and **Confidence intervals for fit values**.

Further variables to be added to the data file can be specified, as follows.

- ◇ Click on the **Save** tab.
- ◇ In the **Save Variables** area, select all four options by checking their check boxes in the **Save** column: **Predicted Values, Lower Confidence Limits, Upper Confidence Limits** and **Noise Residuals**.

These additional variables will not be used in this activity, but it is important nevertheless to know how to obtain them.

Finally, specify the predicted values to be calculated, as follows.

- ◇ Click on the **Options** tab.
- ◇ Select **First case after end of estimation period through a specified date**.
- ◇ Under **Date**, enter 2010 in the **Year** field and 1 in the **Quarter** field.
- ◇ The **Confidence Interval Width (%)** field is used to specify the prediction limits for the forecasted values. Leave this value set to 95, for 95% prediction limits. Leave all other settings as they are.

This completes the set-up.

- ◇ Click on **OK** to run the model.

Most of the output in the **Viewer** window is the same as that obtained when fitting the model without forecasting future values. The main difference is the last item, which is a multiple time plot similar to that shown in Figure 8.10.

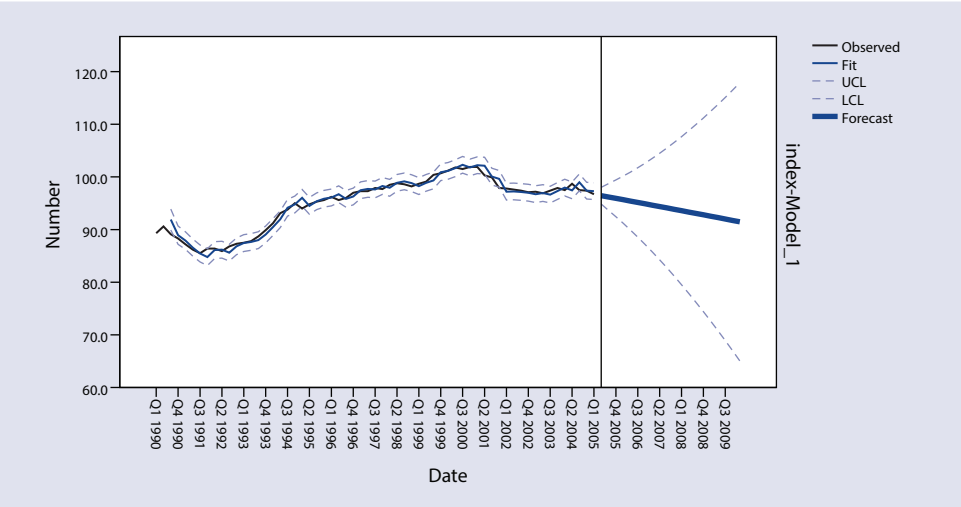


Figure 8.10 Observed values, forecasted values and prediction limits

This shows the data and the 1-step ahead forecasts for the period from the first quarter of 1990 to the first quarter of 2005, along with the forecasts from the second quarter of 2005 until the first quarter of 2010. The vertical line at quarter two of 2005 indicates the start of the forecasting period. Also shown are the 95% prediction intervals, denoted UCL (for ‘Upper Confidence Limit’) and LCL (for ‘Lower Confidence Limit’).

For the period in which data are available — that is, up to the first quarter of 2005 — the prediction intervals for the 1-step ahead forecasts are quite narrow. After the first quarter of 2005, the forecasts are increasingly uncertain. This is reflected by the dramatic widening of the prediction intervals.

Now look in the **Data View** panel of the **Data Editor** and scroll down to the end of the file. You will find that the variables `YEAR_`, `QUARTER_` and `DATE_` have been extended up to the first quarter of 2010. Values of `Predicted_index_Model_1` have been calculated until this date. These are the forecasts for the variable `index`, based on the fitted $ARIMA(0, 2, 1)$ model.

Notice also that variables `LCL_index_Model_1` and `UCL_index_Model_1` have been created. These are the 95% prediction limits for the forecasted values (they are *not* confidence limits). At each time point, the interval (lower, upper) at that time point defines a 95% prediction interval for the forecasted value at that time point. For example, the forecasted value for the first quarter of 2010 is 91.5 (to one decimal place), with 95% prediction interval (65.1, 117.9). The final new variable in the data file is `NResidual_index_Model_1`, which is the 1-step ahead forecast error of the differenced series.

The added variables, after ‘tidying up’, can be used to obtain further plots and to check the adequacy of the model.

The results can be summarized as follows.

The UK index of production is forecasted to fall gradually from the value 96.7, which was recorded in the first quarter of 2005. The forecasted value for the first quarter of 2010 is 91.5, though there is considerable uncertainty around this forecast (95% prediction limits 65.1, 117.9).

Computer Activity 8.10 will give you some practice at obtaining and interpreting forecasts and prediction intervals.

Computer Activity 8.10 *Forecasting the numbers of airline passengers*

The SPSS data file **airline5.sav** contains the time series of annual numbers of airline passengers in the UK. Open the file now. In Computer Activity 8.8, you fitted an ARIMA(0, 2, 1) model to the logarithms of the passenger numbers.

- (a) Obtain the **Time Series Modeler** dialogue box and click on **Reset** to restore default settings. Fit the model, and obtain a multiple time plot of the data, together with forecasts and 95% prediction limits for the annual numbers of passengers up to 2010.
 - (b) Interpret your findings: describe the projected trend between 1999 and 2010, the forecast for 2010, and the uncertainty in the projected numbers of passengers. Make sure you round your results appropriately.
 - (c) Discuss briefly the validity of your forecasts in the light of any events since 1999 which you think might be relevant.
-

Summary of Chapter 8

In this chapter, you have learned how to obtain the correlogram and the partial correlogram for a time series — after applying a log transformation or differencing (or both) to produce stationarity, if necessary. You have used these plots to select possible ARIMA models for time series. You have also learned how to fit an ARIMA model and check its adequacy, and how to obtain forecasts based on an ARIMA model and prediction intervals for the forecasts.

Computer Exercises on Book 2

Computer Exercise 1 Sea ice thickness in Resolute Bay

The SPSS data file **ice.sav** contains data on the maximum thickness (in cm) of sea ice in Resolute Bay, Canada, in the spring of each year between 1960 and 1997. There are three variables: **thickness**, **YEAR_** and **DATE_**.

These data were obtained in July 2005 from the website of the Unaami Data Collection <http://www.unaami.noaa.gov>.

- Obtain a time plot of **thickness**, and describe the time series briefly. Is an additive model appropriate for these data? Explain your answer.
- Obtain a simple centred moving average of order 9, and a multiple time plot of the data and the moving average. Describe the underlying trend in the data.
- Use simple exponential smoothing to forecast the maximum ice thickness for the spring of 1998. Obtain the optimal value of the smoothing parameter and the *RMSE*. Comment on the value of the smoothing parameter. Calculate the *SSE* and obtain the forecast for the spring of 1998.
- Obtain and interpret the time plot, the histogram with superimposed normal curve and the correlogram for the forecast errors, and carry out the Ljung–Box test for lags 1 to 20. Is it reasonable to assume that the forecast errors are white noise?
- Calculate an approximate 95% prediction interval for the forecasted maximum ice thickness in the spring of 1998, and summarize your results.

Computer Exercise 2 UK claimant count

The claimant count is one of several measures used to evaluate the number of persons seeking work in the UK. The variable **claimants** in the SPSS data file **claimants.sav** contains the monthly claimant count, in thousands, between January 1998 and June 2005.

These data were obtained from the National Statistics website <http://www.statistics.gov.uk> in July 2005.

- Obtain a time plot of **claimants** by **DATE_**, with time axis labels only for January and July of each year. As far as you can tell from the time plot, is an additive model appropriate for this time series?
- Assuming that an additive model is appropriate, estimate the seasonal factors of the time series. Interpret the seasonal factors. Suggest a reason why the seasonal factors for October to December are negative.
- Estimate the trend by smoothing the seasonally adjusted series with a simple centred moving average of order 11. Hence estimate the irregular component. Obtain a time plot of the irregular component.
- The Holt–Winters method is to be used to obtain forecasts using the unadjusted variable **claimants** for the twelve months from July 2005 to June 2006. Obtain the optimal values of the smoothing parameters. Comment briefly on the values of these parameters.
- Obtain a time plot of the observed and forecasted numbers of claimants. Briefly summarize how the predicted claimant count changes between June 2005 and June 2006.

Computer Exercise 3 *The river Rhymney at Llanedeyrn*

The SPSS data file **flow.sav** contains data on the monthly average flow of the river Rhymney recorded at Llanedeyrn, Wales, between January 1973 and December 2003.

The variable **flow** contains the time series of monthly average flows (in $\text{m}^3 \text{s}^{-1}$), **logflow** contains the natural logarithms of the values in **flow**, and **saflow** contains the seasonally adjusted values of the time series in **logflow**.

These data were obtained in July 2005 from the website of the Centre for Ecology and Hydrology
<http://www.nwl.ac.uk/ih>.

- (a) Obtain a time plot of **saflow**. Is the time series stationary in mean? Is it stationary in variance? Explain your answers.
- (b) Obtain the correlogram and the partial correlogram for **saflow**, for lags 1 to 20. Briefly describe the main features of these correlograms.
- (c) Suggest one or more possible ARIMA models for **saflow**, giving reasons for your suggestions.
- (d) Fit an ARIMA(1,0,0) model and an ARIMA(0,0,1) model to **saflow**. For each model, obtain the *RMSE*. Also obtain the estimated values of the parameters, and hence write down the model equations for each of the models.
- (e) Check the adequacy of the fit of the model in part (d) that had the smaller *RMSE*.
- (f) Use the model with the smaller *RMSE* to forecast the seasonally adjusted log flow for December 2004. Obtain a time plot of the observed and forecasted values, together with 95% prediction limits. Briefly comment on the width of the prediction intervals.

Learning outcomes

You have been working to acquire the following skills in using SPSS.

- ◇ Enter time series data and use SPSS to define dates.
- ◇ Produce time plots and multiple time plots, and edit the time axis labelling.
- ◇ Transform data using power transformations.
- ◇ Calculate simple centred moving averages.
- ◇ Alter the scales on which time plots are drawn.
- ◇ Estimate the seasonal factors of a time series that can be described adequately by an additive decomposition model, and obtain the seasonally adjusted series.
- ◇ Use moving averages to estimate the trend component and the irregular component of a time series that can be described by an additive decomposition model.
- ◇ Obtain forecasts using simple exponential smoothing, Holt's exponential smoothing and Holt–Winters exponential smoothing.
- ◇ Obtain optimal values for the parameters of an exponential smoothing method.
- ◇ Obtain the *RMSE* and calculate the *SSE* from it.
- ◇ Obtain the correlogram and the partial correlogram for a time series.
- ◇ Carry out the Ljung–Box test for zero autocorrelation.
- ◇ Obtain a histogram with a normal curve superimposed.
- ◇ Produce a time plot of a time series after a transformation or differencing (or both).
- ◇ Fit an ARIMA model, and obtain forecasts using the model and prediction intervals for the forecasts.

Solutions to Computer Activities

Solution 1.3

In the **Define Dates** dialogue box, select the option **Years, quarters** in the **Cases Are** area. Enter 1991 in the **Year** field and 1 in the **Quarter** field in the **First Case Is** area.

The method is described in Computer Activity 1.2.

Solution 1.6

- (a) Instructions for obtaining a time plot are given in Computer Activity 1.4. The time plot required is shown in Figure S.1(a).

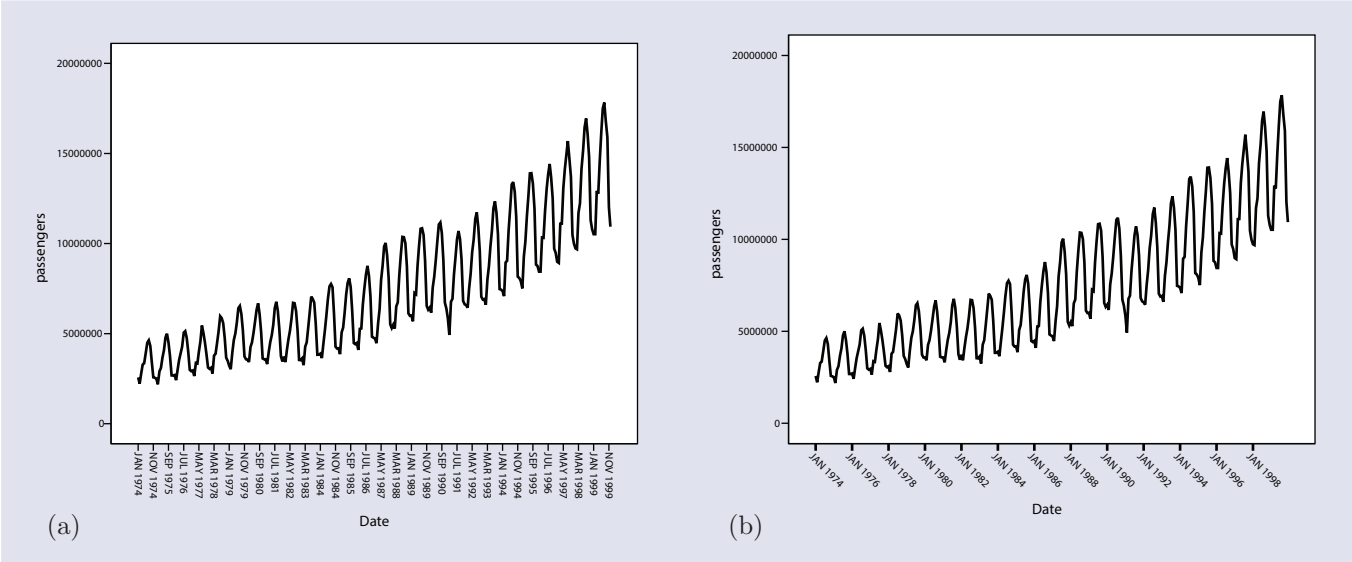


Figure S.1 Monthly numbers of passengers: (a) the default time plot (b) an edited time plot

The number of airline passengers increased between 1974 and 1999: the trend is curved upwards. There is marked cyclic variation. The size of the cyclic fluctuations increases with the level of the series.

- (b) The use of the **Chart Editor** to edit the labelling on the time axis is described in Computer Activity 1.4. Labels will be included only for each January if you enter 11 in the **Ticks skipped between labels** field. The time plot produced by SPSS is shown in Figure S.1(b) above. Notice that SPSS has displayed labels for January every two years, not every year as requested.

Labelling every other January on the time axis makes it easier to see that the lowest point of each cycle occurs in winter. Thus the cycle is annual and the series is seasonal.

(c) The time plot required is shown in Figure S.2.

The method is described in Computer Activity 1.5.

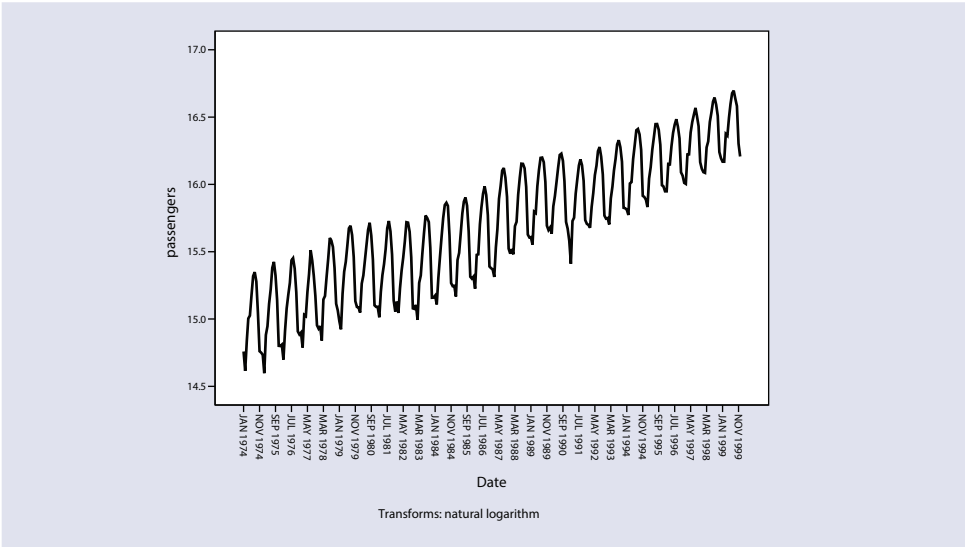


Figure S.2 Monthly numbers of passengers, logarithms

(d) An additive model is not appropriate for the original time series because the size of the seasonal fluctuations increases with the level of the series, as you found in part (a). If a multiplicative model were suitable, then an additive model would be appropriate for the time series of logarithms. The time plot in Figure S.2 shows that this is not the case: the size of the seasonal fluctuations for the time series of logarithms decreases as the level increases. Thus neither an additive model nor a multiplicative model is appropriate for the time series of monthly numbers of passengers.

Solution 1.8

- (a) The use of the **Compute Variable** dialogue box to transform a variable using a power function is described in Computer Activity 1.7.
- (b) The time plot required is shown in Figure S.3.

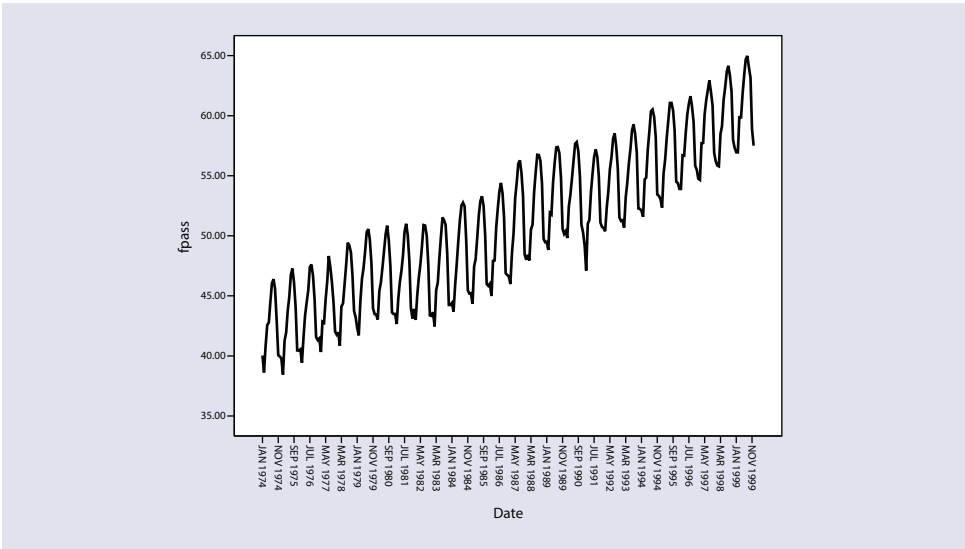


Figure S.3 Monthly numbers of passengers, fourth roots

The size of the seasonal fluctuations is roughly the same whatever the level of the series. This suggests that the time series of fourth roots of passenger numbers might be described adequately by an additive model.

Solution 2.3

(a) The time plot of yield is shown in Figure S.4.

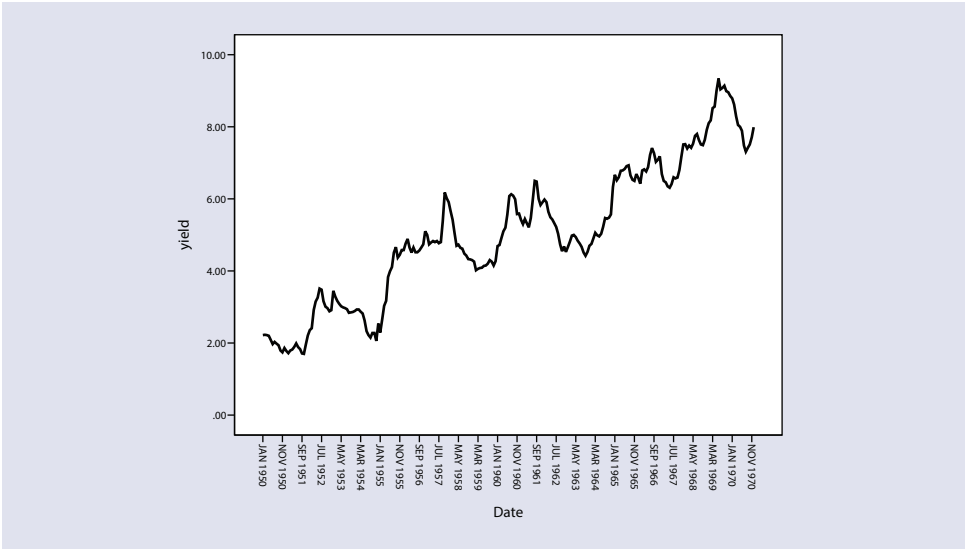


Figure S.4 Time plot of monthly percentage yields

There is an upward trend in the monthly percentage yield. The irregular fluctuations do not appear to increase in size as the level increases.

(b) The method for calculating a simple centred moving average is described in Computer Activity 2.1. Type **ma9** in the **Name** field of the **Create Time Series** dialogue box, and enter **9** in the **Span** field. Remember to delete the label for **ma9**. The time plot of **ma9** with labels on the time axis for each January is shown in Figure S.5.

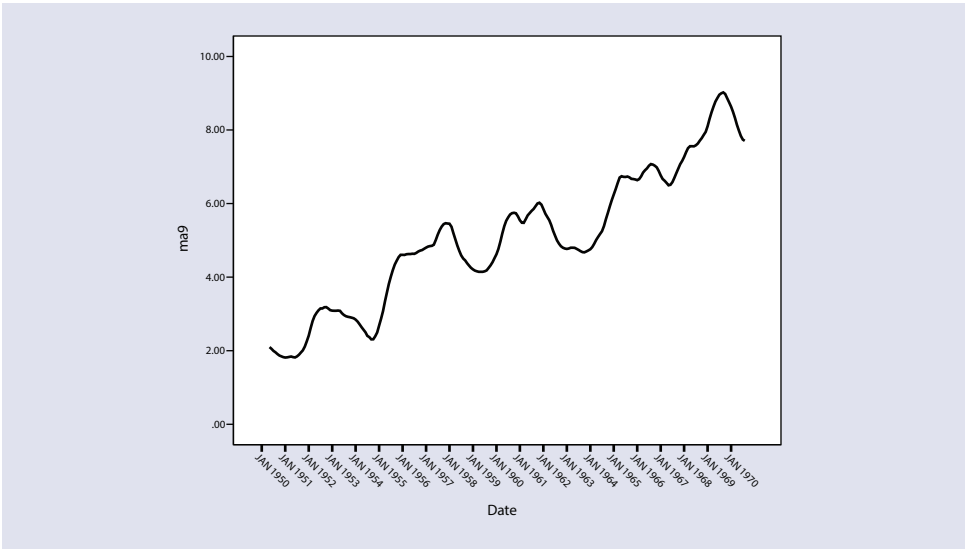


Figure S.5 Time plot of moving average of order 9

Peaks occurred, roughly, in 1952, 1957, 1961, 1966 and 1969. Note that it is easier to identify the peak years if you edit the labelling of the time axis so that the labels occur at one-year intervals: the method is described in Computer Activity 1.4.

- (c) The minimum and maximum values on the vertical scale are 0 and 10, respectively. These may be found using the method described in Computer Activity 2.2.
- (d) The new variable `ma21` can be created using the method described in Computer Activity 2.1. Type `ma21` in the **Name** field, and `21` in the **Span** field. Remember to delete the label of the newly created variable `ma21`.

The vertical scale on the time plot of `ma21` can be changed using the method described in Computer Activity 2.2. The values obtained in part (c) must be typed in the **Minimum** and **Maximum** fields. The time plot of `ma21` drawn on the scale used for the time plot of `ma9` is shown in Figure S.6.

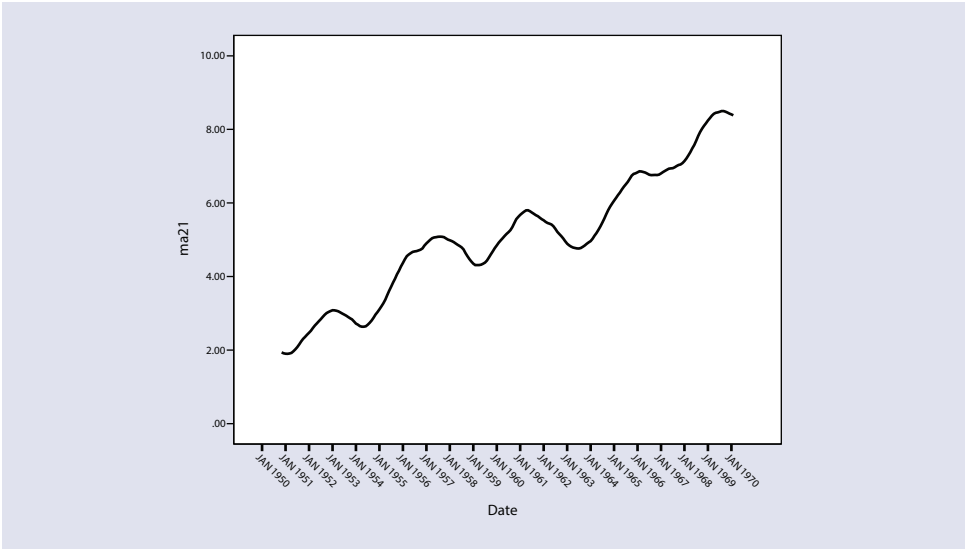


Figure S.6 Time plot of `ma21` drawn using the same scale as `ma9`

- (e) The moving average of order 21 has smoothed out the bumps even more than the moving average of order 9: for example, it is no longer clear that there is a peak in 1966. This time series might be over-smoothed: the moving average of order 9 might provide the better balance between under-smoothing and over-smoothing. On the other hand, the fluctuations shown by `ma9`, that have been ironed out by `ma21`, may be irrelevant noise; in this case, `ma21` would be the better choice. Note that in answering questions such as this one, you should attempt to provide a coherent argument supporting your choice. The choice you make is less important than the argument you provide for it.

Solution 3.4

- (a) Obtaining and editing a time plot is described in Computer Activity 1.4. To obtain the required spacing of time point labels, you should type 23 in the **Ticks skipped between labels** field in the **Labels & Ticks** panel of the **Properties** dialogue box. The time plot required is shown in Figure S.7.

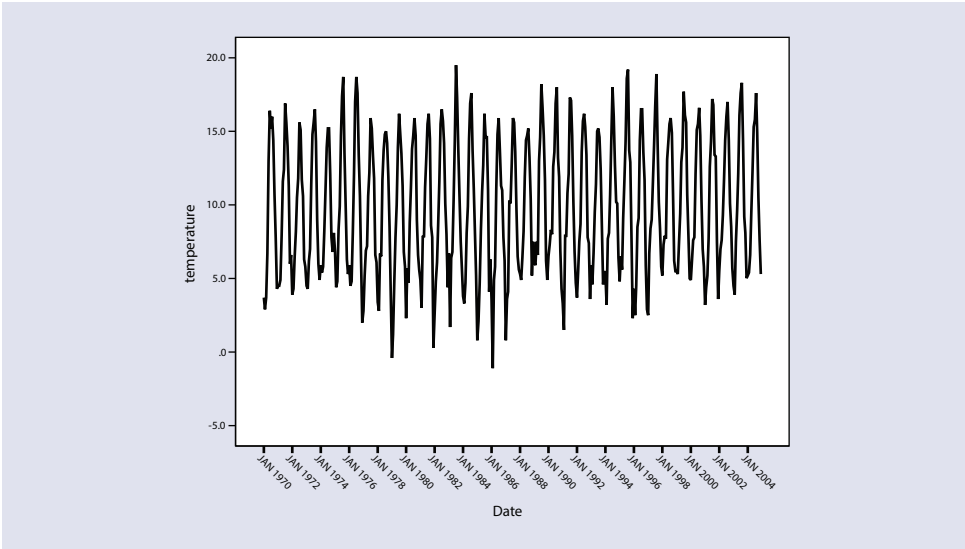


Figure S.7 Time plot of monthly average temperatures

- The series fluctuates seasonally around an average of about 10°C. This time plot does not provide any evidence of a change in the level (that is, of a trend) in the series between 1970 and 2004.
- (b) Estimating the seasonal component is described in Computer Activity 3.1. The estimated seasonal factors are given in the **Seasonal Factors** output table, which is reproduced in the margin. (Note that SPSS cuts off the final ‘e’ from the `temperature` label.)

The seasonal maximum occurs in July (6.6891 in Period 7). The seasonal minimum occurs in January (−5.5639 in Period 1).

- (c) It is good practice to rename the seasonally adjusted series `adjusted` and delete its label. The simple centred moving average of order 35 for `adjusted` is obtained using the method described in Computer Activity 2.1. For clarity, the variable name `trend` should be used for the moving average, and its label deleted to avoid clutter.

Seasonal Factors	
Series Name: temperatur	
Period	Seasonal Factor
1	-5.5639
2	-5.4666
3	-3.4898
4	-1.6343
5	1.5986
6	4.3623
7	6.6891
8	6.5107
9	3.9541
10	.6472
11	-2.8110
12	-4.7963

A time plot of the estimated trend component is shown in Figure S.8.

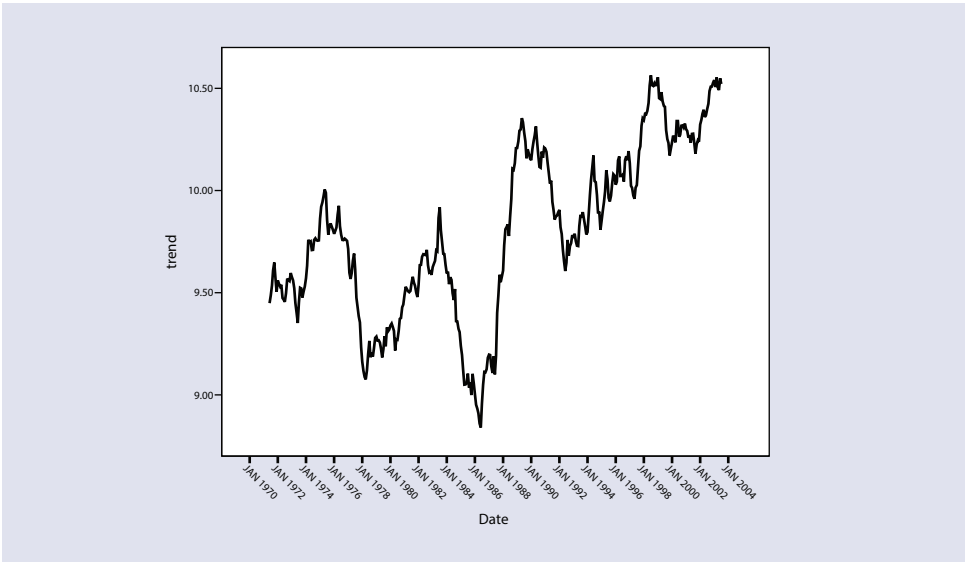


Figure S.8 Estimated trend component

This time plot is quite jagged and complicated to interpret. There appears to be a cycle of period about 8 years, and perhaps a rising trend since the late 1980s. However, the main conclusion is that in view of the variability in these data, a longer time series is required to get a clear picture of the underlying trend.

- (d) The irregular component is obtained by subtracting the trend component from the seasonally adjusted series. A time plot of the irregular component, which has been named `irregular`, is shown in Figure S.9.

The method is described in Computer Activity 3.3.

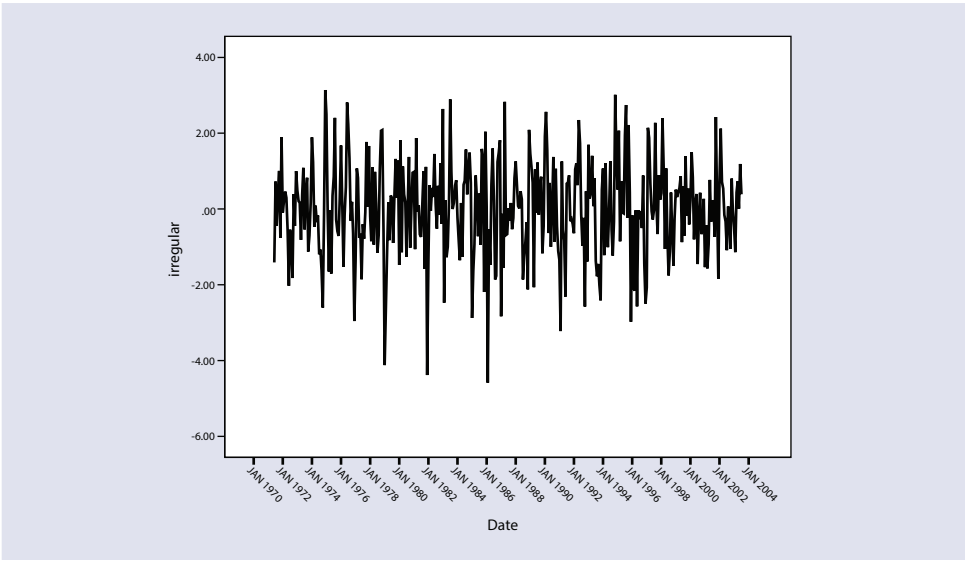


Figure S.9 Estimated irregular component

From part (b), the seasonal variation spans about 12°C: the seasonal low is -5.5, the seasonal high is +6.7, compared to the annual averages. The irregular component seldom fluctuates by more than 4°C (see Figure S.9), while the change in level of the time series between 1970 and 2004 is of the order of 1°C (see Figure S.8). So the seasonal component dominates the time series.

The SPSS data file `temperature4.sav` contains the estimated components if you wish to check your results.

Solution 4.2

(a) A time plot of the annual precipitation is shown in Figure S.10.

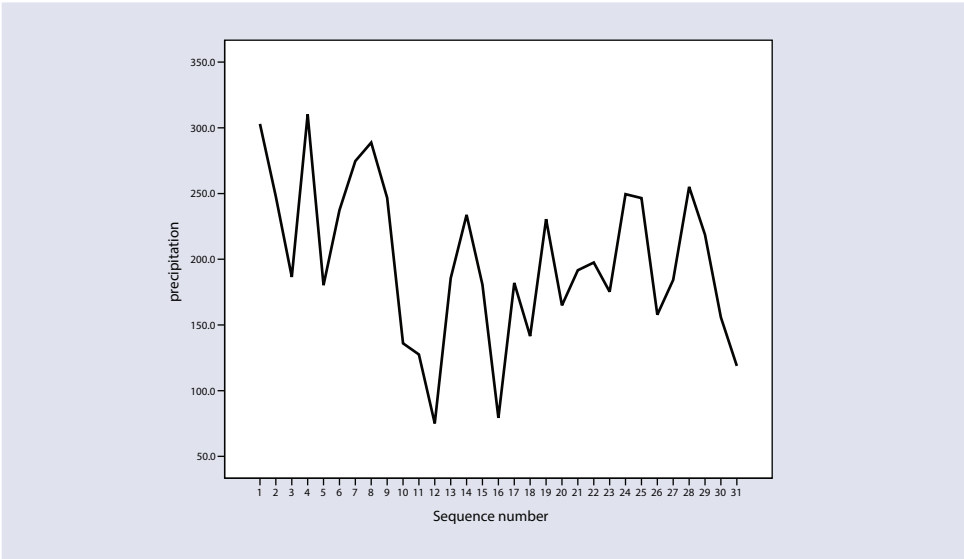


Figure S.10 Time plot of annual precipitation

The data are annual (hence non-seasonal) and there is no clear trend. The size of the fluctuations does not appear to vary. So the simple exponential smoothing method is appropriate.

(b) The optimal smoothing parameter is $\alpha = 0.251$. The $RMSE$ is 59.854. Since there are $n = 31$ observations and $k = 1$ smoothing parameters,

$$SSE = (31 - 1) \times 59.854^2 \simeq 107\,480.$$

Obtaining forecasts is described in Computer Activity 4.1.

(c) A multiple time plot is shown in Figure S.11.

Obtaining a multiple time plot is described in Computer Activity 4.1.

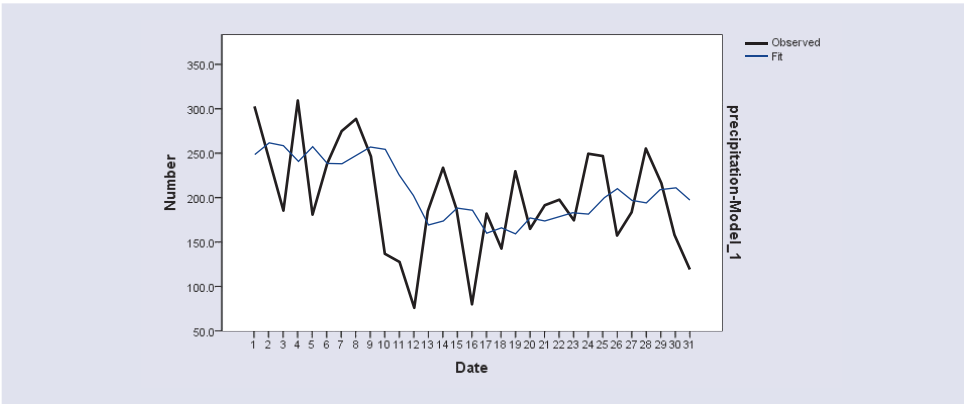


Figure S.11 Observed and forecasted values

The forecasts are much smoother than the observations, reflecting the relatively low value of α . They track the changes in the trend with a slight delay, notably the drop in the level between time points 8 and 12.

Solution 4.4

- (a) The forecasts for the years 1998/9 to 2001/2 are all equal to 177.9 to one decimal place.
- (b) A multiple time plot is shown in Figure S.12.

Obtaining forecasts is described in Computer Activity 4.3.

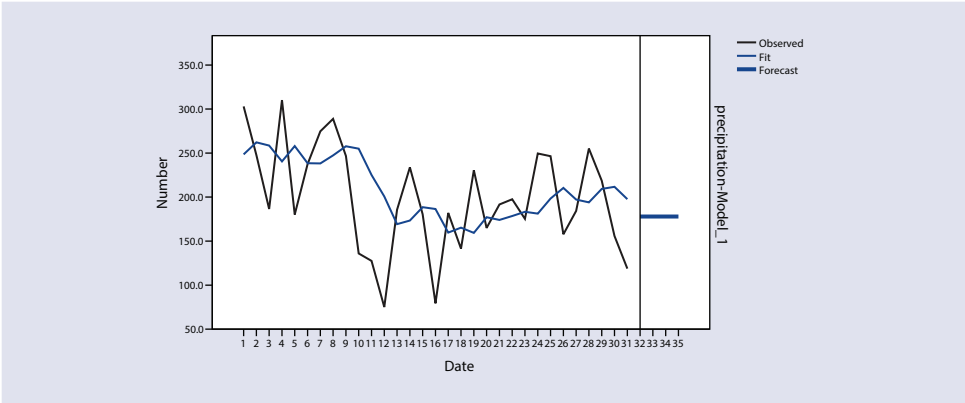


Figure S.12 Observed and forecasted values

The forecasts for the years 1998/9 to 2001/2, corresponding to time points 32 to 35, are represented as a horizontal line, because simple exponential smoothing assumes no trend.

Solution 5.2

- (a) A time plot is shown in Figure S.13.

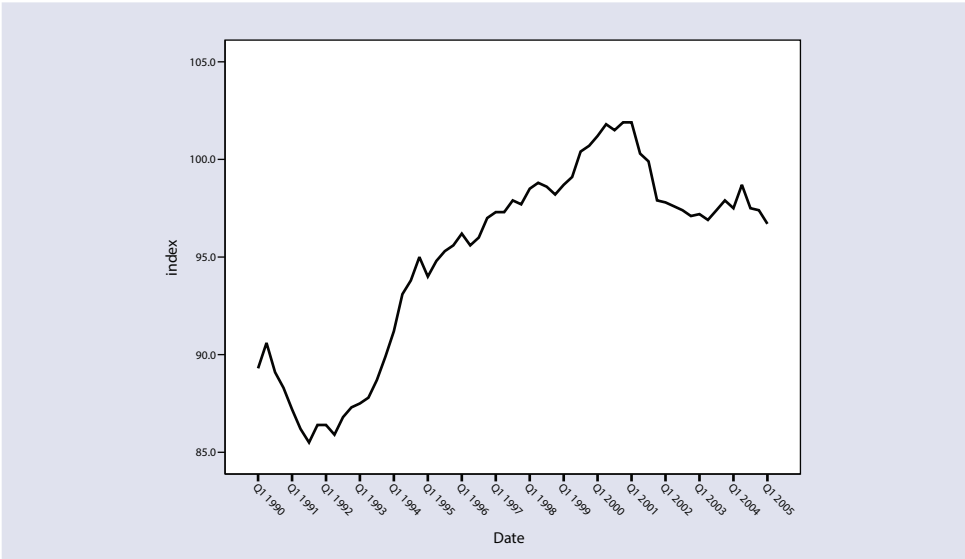


Figure S.13 Time plot of the UK index of production

The production index varies in a rather complicated pattern involving alternating periods of growth and decline. It is certainly not constant (on average) over time, so the simple exponential smoothing method is inappropriate. Holt’s exponential smoothing method should work within each growth and contraction period, but will not predict when the trend changes. But then, it’s probably the case that no other statistical forecasting method will either!

- (b) The optimal values of the smoothing parameters are $\alpha = 0.887$ and $\gamma = 0.370$. The corresponding $RMSE$ is 0.777.

There are $n = 61$ observations and $k = 2$ smoothing parameters. So

$$SSE = (61 - 2) \times 0.777^2 \simeq 35.6.$$

- (c) The multiple time plot is shown in Figure S.14.

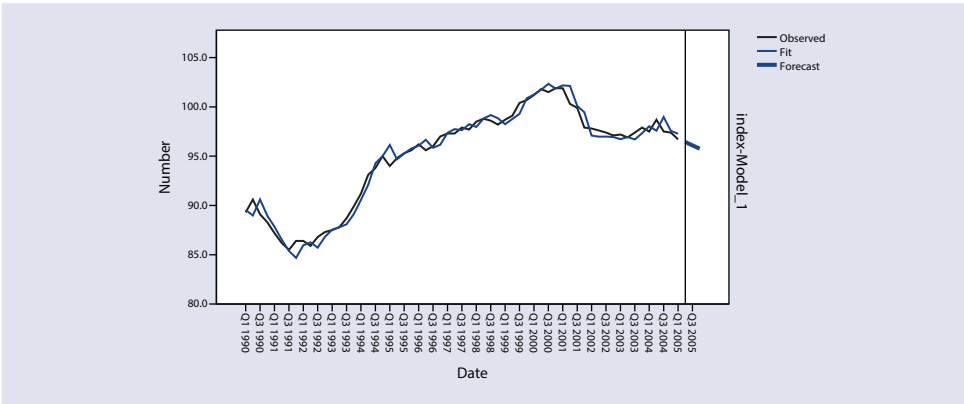


Figure S.14 Observed and forecasted values of the index

The forecasts are close to the observed values. A decrease in the production index is forecasted over the last three quarters of 2005. The production index is forecasted to drop to 95.8 in the last quarter of 2005.

Solution 5.4

- (a) A time plot is shown in Figure S.15.

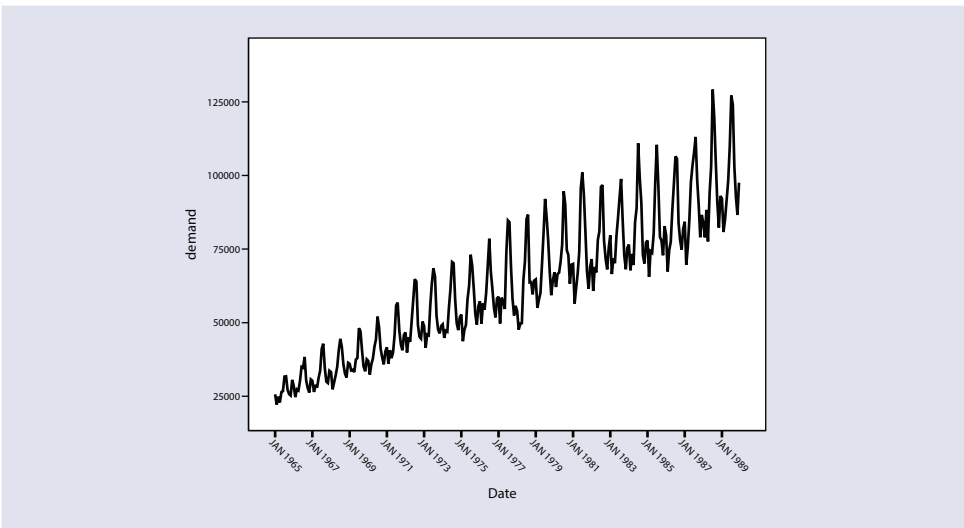


Figure S.15 Time plot of demand

The magnitude of the seasonal fluctuations (and also the irregular variation) increases with the level of the series. So an additive model is not appropriate for this time series.

(b) A time plot of logdemand is shown in Figure S.16.

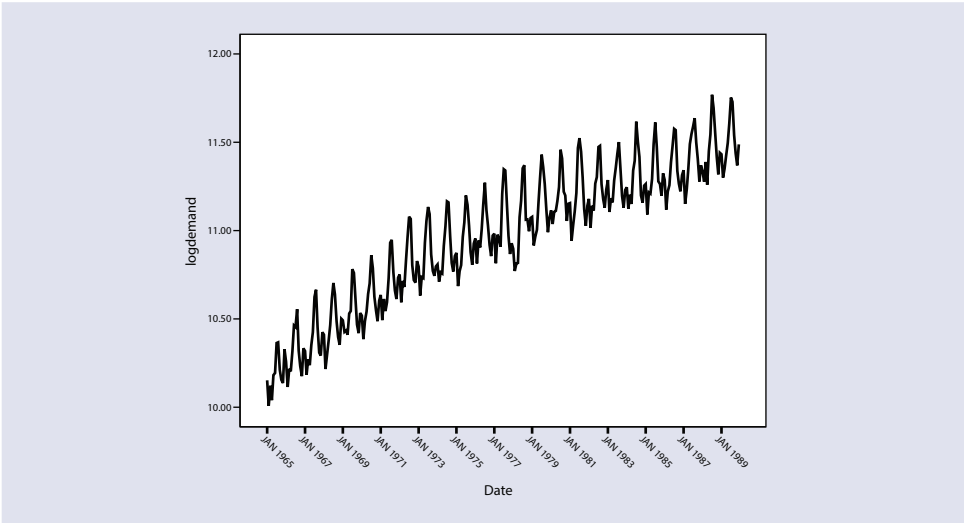


Figure S.16 Time plot of logdemand

The magnitude of the seasonal fluctuations, and of the irregular variation, is roughly the same whatever the level. Hence an additive model is suitable for representing this time series.

- (c) The values required are in the **Exponential Smoothing Model Parameters** output table, which is reproduced below.
- The method is described in Computer Activity 5.3.

Exponential Smoothing Model Parameters						
Model			Estimate	SE	t	Sig.
logdemand-Model_1	No Transformation	Alpha (Level)	.407	.047	8.712	.000
		Gamma (Trend)	4.249E-006	.005	.001	.999
		Delta (Season)	2.922E-005	.044	.001	.999

The optimal parameter values are $\alpha = 0.407$, $\gamma = 0.000$ and $\delta = 0.000$ (to three decimal places). The corresponding *RMSE* (from the **Model Statistics** output table) is 0.045. There are $n = 300$ observations and $k = 3$ smoothing parameters. Hence

$$SSE = (300 - 3) \times 0.045^2 \simeq 0.60.$$

- (d) A multiple time plot is shown in Figure S.17.

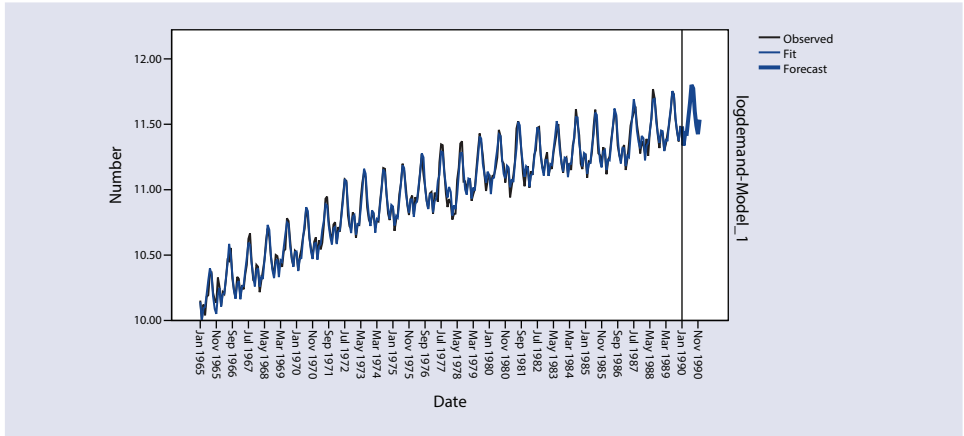


Figure S.17 Observed and forecasted values of logdemand

The 1-step ahead forecasts match the observed values very closely up to December 1989. The Holt–Winters forecasts take account of both trend and seasonality.

Looking at the forecasted values in the **Forecast** output table, the forecasted value for the logarithm of the monthly electricity demand in 1990 is lowest in February (11.34) and highest in July (11.80).

Solution 6.2

(a) The correlogram required is shown in Figure S.18.

The method is described in Computer Activity 8.1.

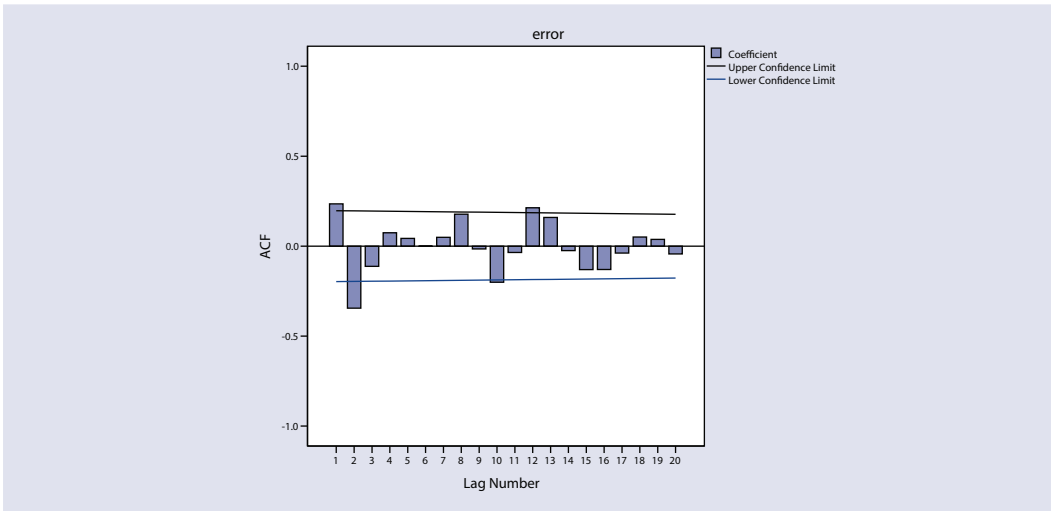


Figure S.18 Correlogram for forecast errors

(b) The sample autocorrelations at lags 1 and 2 clearly exceed the significance bounds. The sample autocorrelations at lags 10 and 12 just do so as well.

The values of these sample autocorrelations are given in the **Autocorrelations** output table. This is displayed in the **Viewer** window immediately above the correlogram.

Scroll up the output to view this table.

The lags and sample autocorrelations are in the first two columns. The sample autocorrelations at lags 1 and 2 are 0.235 and -0.344 , respectively. The autocorrelations at lags 10 and 12 are -0.200 and 0.213 . You should not make too much of the autocorrelations at lags 10 and 12: recall that ‘it is all too easy to over-interpret correlograms’. But the autocorrelations at lags 1 and 2 are substantial, and suggest a clear departure from the white noise model.

(c) The value of the Ljung–Box test statistic and the p value for lags 1 to 20 can be read from the bottom row of the fourth and sixth columns of the **Autocorrelations** output table. Thus the test statistic is 41.958, and $p = 0.003$. This provides strong evidence against the null hypothesis that all autocorrelations at lags 1 to 20 are zero. Hence there is strong evidence that there is autocorrelation at some lag between 1 and 20.

(d) Overall, there is strong evidence against the white noise model for the forecast errors. The correlogram suggests that there is non-zero autocorrelation at lags 1 and 2. This suggests that the simple exponential smoothing method can be improved upon.

Solution 6.4

- (a) The use of the **Chart Editor** to edit the labels on the time axis is described in Computer Activity 1.4. To position the time axis labels at January of each year, enter 11 in the **Ticks skipped between labels** field in the **Labels & Ticks** panel of the **Properties** dialogue box. To obtain the time plot in Figure S.19, **Diagonal** was chosen from the **Label orientation** drop-down list.

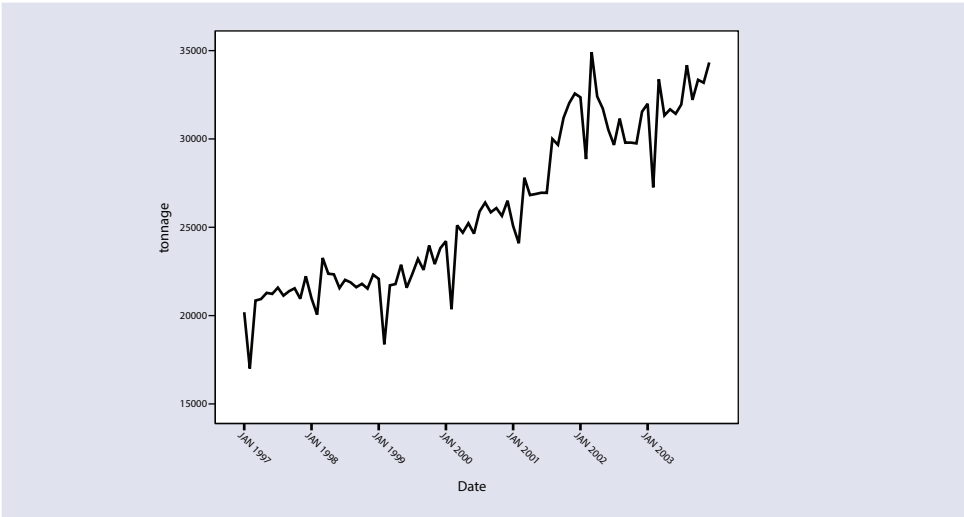


Figure S.19 Time plot of tonnage

The first feature to note is the upward trend between 1997 and 2003. Secondly, there are regularly-spaced downward spikes. Their position relative to the labels on the time axis shows that they are seasonal. The size of the seasonal spikes and the size of the irregular fluctuations do not vary with the level of the series. So an additive model seems to be appropriate for this time series.

Note: In fact, the annual spikes occur in February. This might be due to the fact that February is a shorter month, and perhaps also to shipping being affected by the celebration of the Chinese New Year, which often occurs in February.

- (b) The implementation of the Holt–Winters method in SPSS is described in Computer Activity 5.3. The forecasted value for January 2004 is 34 025 (rounded to the nearest thousand tonnes).
- (c) The optimal parameter values are $\alpha = 0.700$, $\gamma = 8.445\text{E-}0.006 \simeq 0.000$ and $\delta = 0.000$, and the *RMSE* is 780.749.

There are $n = 84$ observations and $k = 3$ smoothing parameters. So

$$SSE = (84 - 3) \times 780.749^2 = 49\,375\,089.08 \simeq 49\,375\,100.$$

The high value of α means that recent observations have a big impact on the forecast. In contrast, since $\gamma = \delta = 0$, recent observations have no impact on the slope and seasonal factors, which are determined solely by the initial values.

(d) The multiple time plot is shown in Figure S.20.

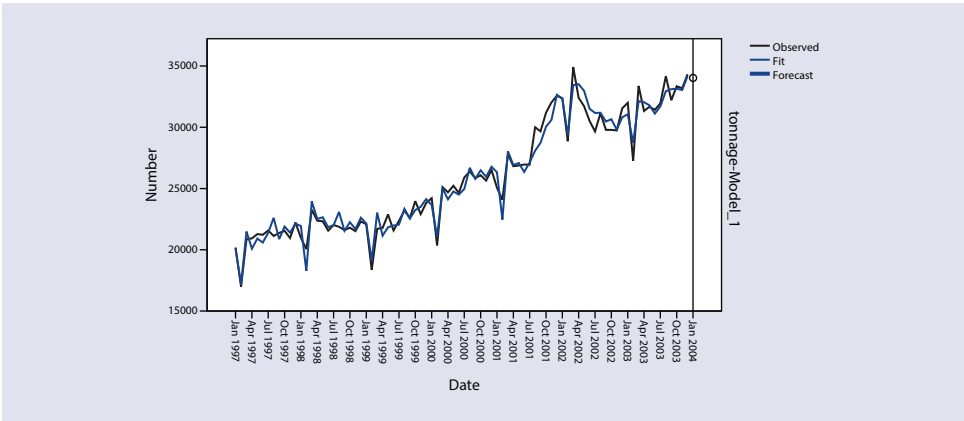


Figure S.20 Observed and forecasted tonnage

The 1-step ahead forecasts fit the observed values well; in particular, the seasonal troughs are correctly accounted for. However, the 1-step ahead forecasts are in-sample forecasts, so may give too optimistic a view of the method’s forecasting ability.

(e) In what follows, the data file has been tidied up. The variable `Predicted_tonnage_Model_1` has been renamed `forecast` and the variable `NResidual_tonnage_Model_1` has been renamed `error`. The labels for these two variables have been deleted. The correlogram is shown in Figure S.21.

Obtaining a correlogram is described in Computer Activity 6.1.

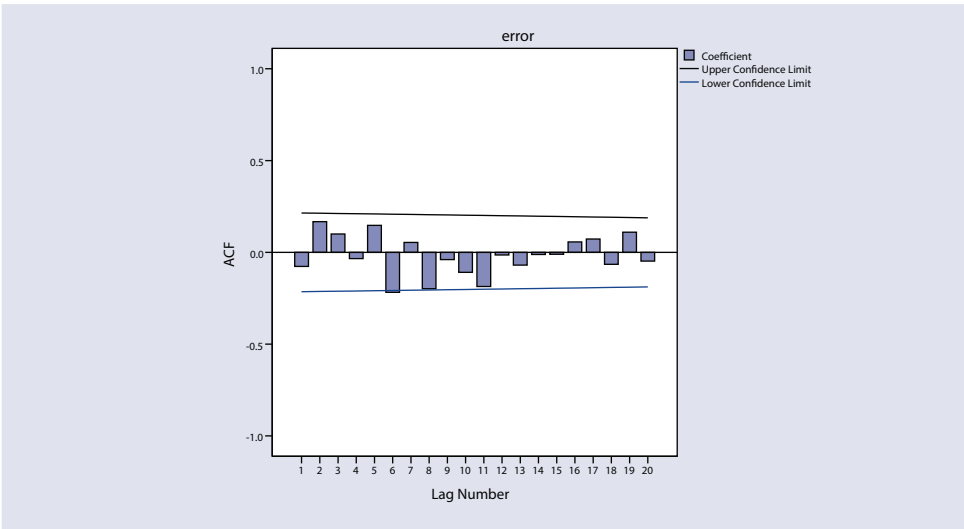


Figure S.21 Correlogram for forecast errors

The sample autocorrelations at lags 6 and 8 just touch the significance bounds, but none of the autocorrelations clearly exceeds the bounds. There does not seem to be much of a pattern in the correlogram.

The value of the Ljung–Box test statistic is 22.622, with p value 0.308. Overall, there is little evidence of non-zero autocorrelation at lags 1 to 20. Hence the correlogram is consistent with the forecast errors being white noise.

The test statistic and p value are given in the `Autocorrelations` output table.

(f) A time plot of the forecast errors is shown in Figure S.22(a).

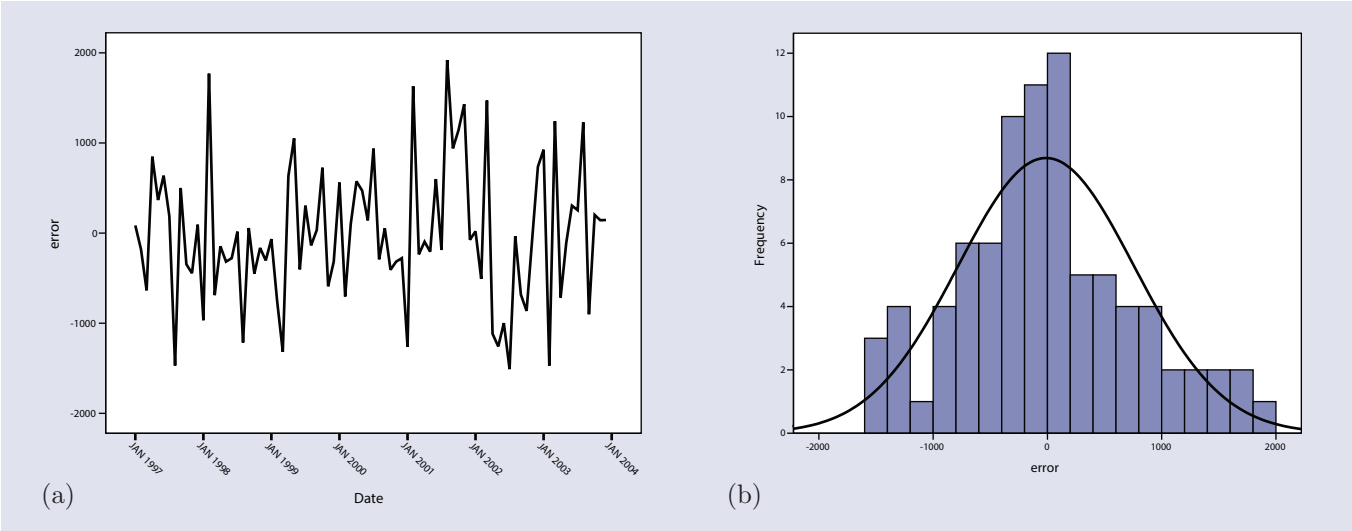


Figure S.22 Forecast errors: (a) time plot (b) histogram

There is no systematic change in level or variance. Thus the forecast errors appear to be distributed around zero with constant variance.

The histogram with superimposed normal curve is shown in Figure S.22(b). The assumption that the forecast errors are normally distributed appears plausible (or, at least, not completely implausible).

- (g) All the assumptions required for the calculation of prediction intervals are satisfied. The forecasted value is 34 025, there are 84 data points (so $n = 84$), and the SSE is 49 375 089.08 $\simeq$ 49 375 100. Hence the 95% prediction limits are as follows:

$$34\,025 - 1.96\sqrt{49\,375\,089.08/84} \simeq 32\,522,$$
$$34\,025 + 1.96\sqrt{49\,375\,089.08/84} \simeq 35\,528.$$

The calculation of prediction intervals is described in Subsection 9.3 of *Book 2*.

Thus the forecasted tonnage (in thousands of tonnes) for January 2004 is 34 025, with 95% prediction interval (32 522, 35 528). Since this contains 32 999, the value actually observed in January 2004, the forecast is reasonable.

The 95% prediction limits calculated by SPSS are in the **Forecast** output table. The lower prediction limit, denoted **LCL**, is 32 471, and the upper prediction limit, denoted **UCL**, is 35 578. Thus, this 95% prediction interval is (32 471, 35 578), which is slightly wider than that calculated according to *Book 2*.

Solution 7.2

- (a) Time plots of the first differences of `ftse` and `logftse` are shown in Figure S.23.

Use the **Sequence Charts** dialogue box, as described in Computer Activity 7.1.

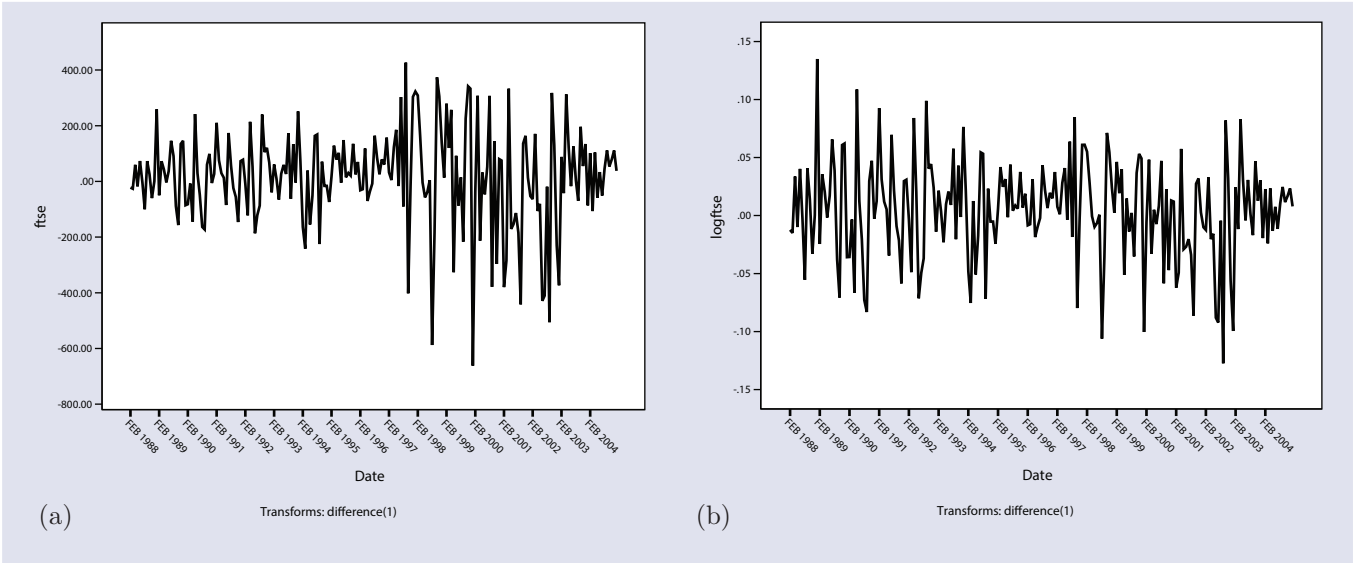


Figure S.23 Time plots of first differences: (a) `ftse` (b) `logftse`

- (b) The first differences of `ftse` are stationary in mean (since there is no systematic variation in level), but not in variance (since the width of the irregular fluctuations suddenly increases about halfway through the time series). The first differences of `logftse` are stationary in mean and in variance: there is no systematic variation in level, and the width of the fluctuations appears to be of constant size.

Solution 7.4

- (a) A time plot of the seasonally adjusted monthly electricity demand is shown in Figure S.24.

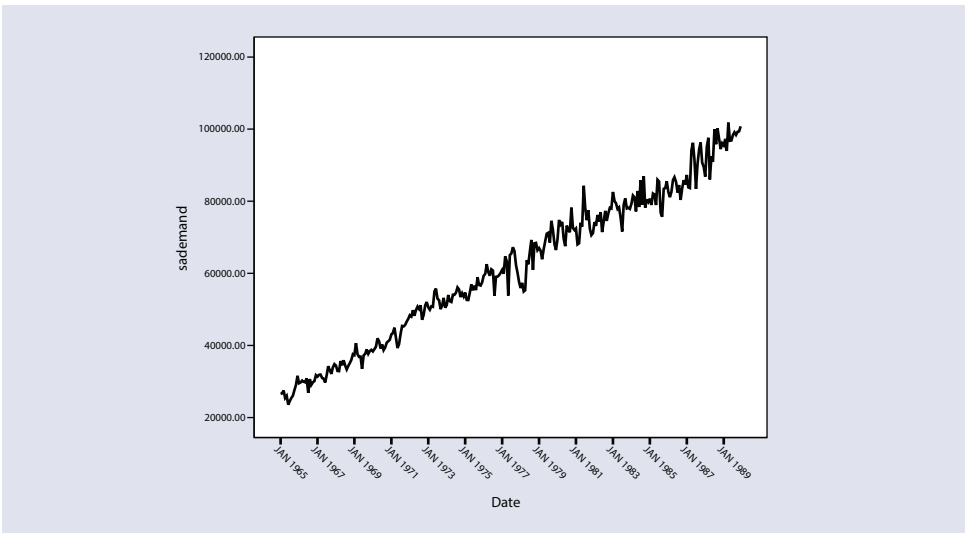


Figure S.24 Time plot of seasonally adjusted demand

The two main features of this time plot are the increasing trend, and the fact that the irregular fluctuations increase in size as the level increases. Thus the time series is not stationary either in mean or in variance.

- (b) In this case, stationarity in variance can be obtained using a log transformation. To obtain stationarity in mean, differencing is required. A time plot of the first differences of the logarithms of `sademand` is shown in Figure S.25.

The method is described in Computer Activity 7.3.

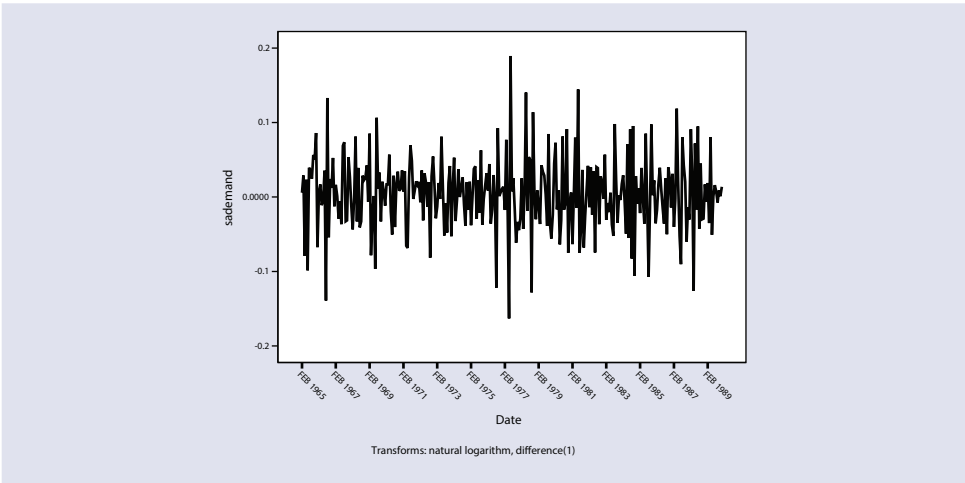


Figure S.25 First differences of logarithms

There is no systematic variation in level, and the size of the irregular fluctuations is roughly constant, so this time series is stationary in both mean and variance. Hence differencing once is sufficient to produce stationarity in mean in this case.

Solution 8.3

- (a) The correlogram and the partial correlogram for lags 1 to 20 are shown in Figure S.26.

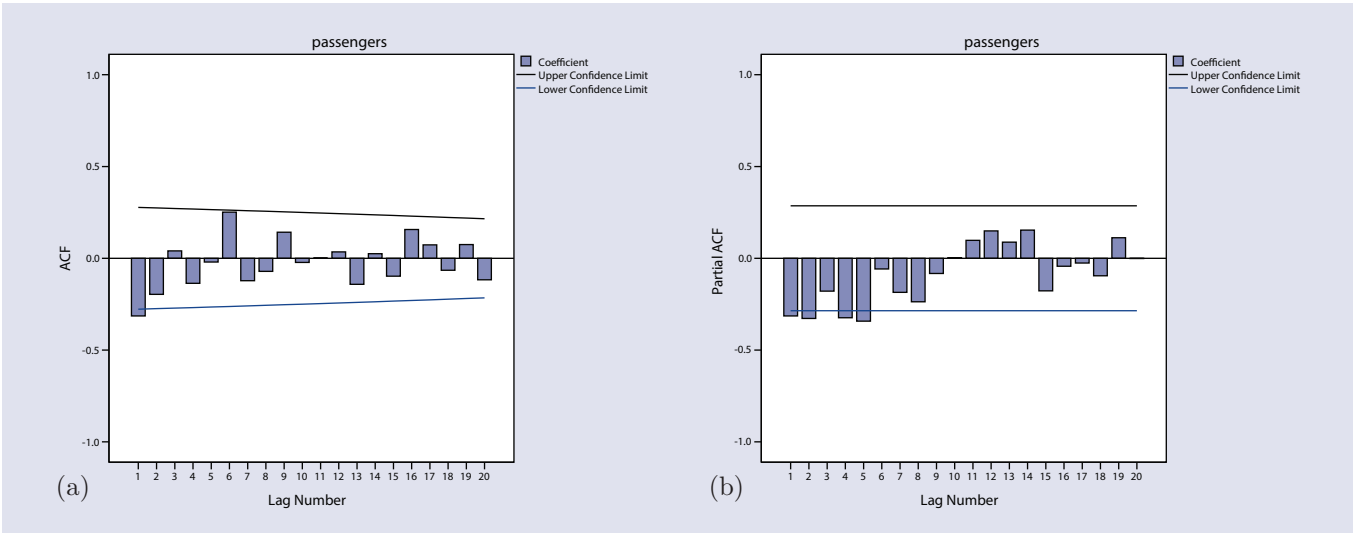


Figure S.26 Second differences of logarithms: (a) correlogram (b) partial correlogram

These may be obtained using the **Autocorrelations** dialogue box, as described in Computer Activity 8.1. **Natural log transform** and **Difference** must be checked, and 2 entered in the **Difference** field. You must also specify that you require 20 lags in the **Autocorrelations: Options** dialogue box.

- (b) The order of differencing d (applied to the logarithm of `passengers`) is 2. The correlogram and partial correlogram suggest a rather weak correlation structure: none of the bars greatly exceeds the significance bounds. In the correlogram, only the autocorrelation at lag 1 exceeds the bounds. In the partial correlogram, the partial autocorrelations are relatively large for the first few lags, then decline in magnitude with increasing lag. These observations suggest that an MA(1) model might be appropriate for the stationary series. Since the autocorrelation at lag 2 is appreciable, a possibility that is not immediately ruled out is an MA(2) model. Another possibility is an ARMA(1, 1) model, since the sample ACF could reasonably be interpreted as ‘tailing off’ rather than abruptly falling to zero after lag 1.
- (c) The principle of parsimony reduces the options to the MA(1) model, since $p + q = 1$ for this model, whereas for the MA(2) model and the ARMA(1, 1) model, $p + q = 2$. Thus a suggested model for the time series of logarithms of the numbers of passengers is ARIMA(0, 2, 1).

Solution 8.6

- (a) The method is as described in Computer Activity 8.4, except that in the **ARIMA Orders** area of the **Time Series Modeler: ARIMA Criteria** dialogue box you should enter 1 in the **Autoregressive (p)** field, 2 in the **Difference (d)** field and 0 in the **Moving Average (q)** field of the **Nonseasonal** column.

The *RMSE* is given in the **Model Statistics** table: $RMSE = 0.840$. The estimates of β_1 and μ are given in the **ARIMA Model Parameters** table: $\hat{\beta}_1 = -0.624$ and $\hat{\mu} = -0.012$.

As there are $n = 61$ observations, $d = 2$ and there are $k = 2$ parameters,

$$SSE = (61 - 2 - 2) \times 0.840^2 \simeq 40.2.$$

- (b) Let Y_t denote the series of second differences. The model formula is

$$Y_t + 0.012 = -0.624(Y_{t-1} + 0.012) + Z_t.$$

The AR(1) model is defined in Subsection 12.1 of *Book 2*.

- (c) Checking the adequacy of a model is described in Computer Activity 8.5. The variables were tidied up in the **Variable View** panel, as follows: the forecasts and errors were renamed `forecast2` and `error2` and their labels were deleted.

A multiple time plot of the observed values and the 1-step ahead forecasts is shown in Figure S.27.

A similar plot could have been obtained from the **Time Series Modeler** dialogue box.

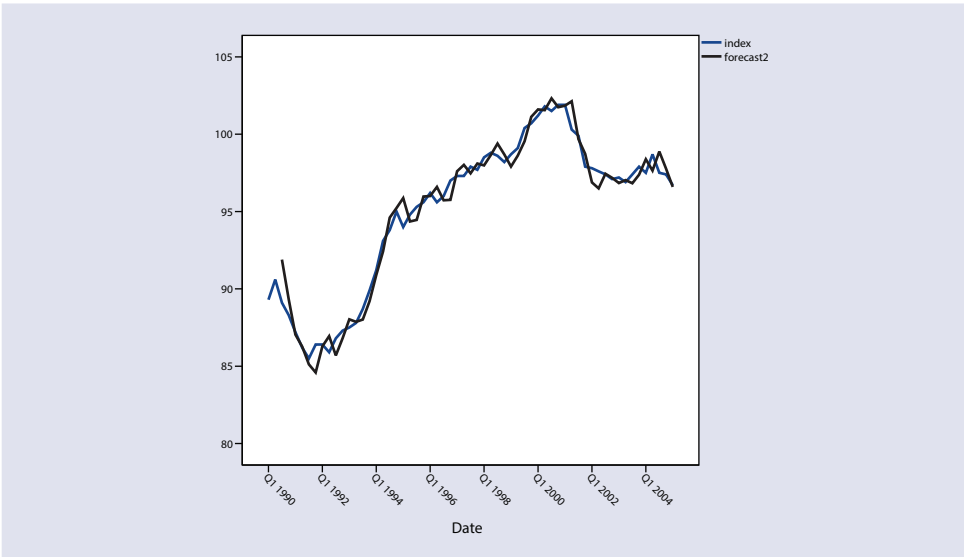


Figure S.27 Observed and forecasted values

This time plot shows that the model gives a reasonably good fit to the data.

A time plot of the forecast errors is shown in Figure S.28(a), and a histogram of the errors with a normal curve superimposed is shown in Figure S.28(b).

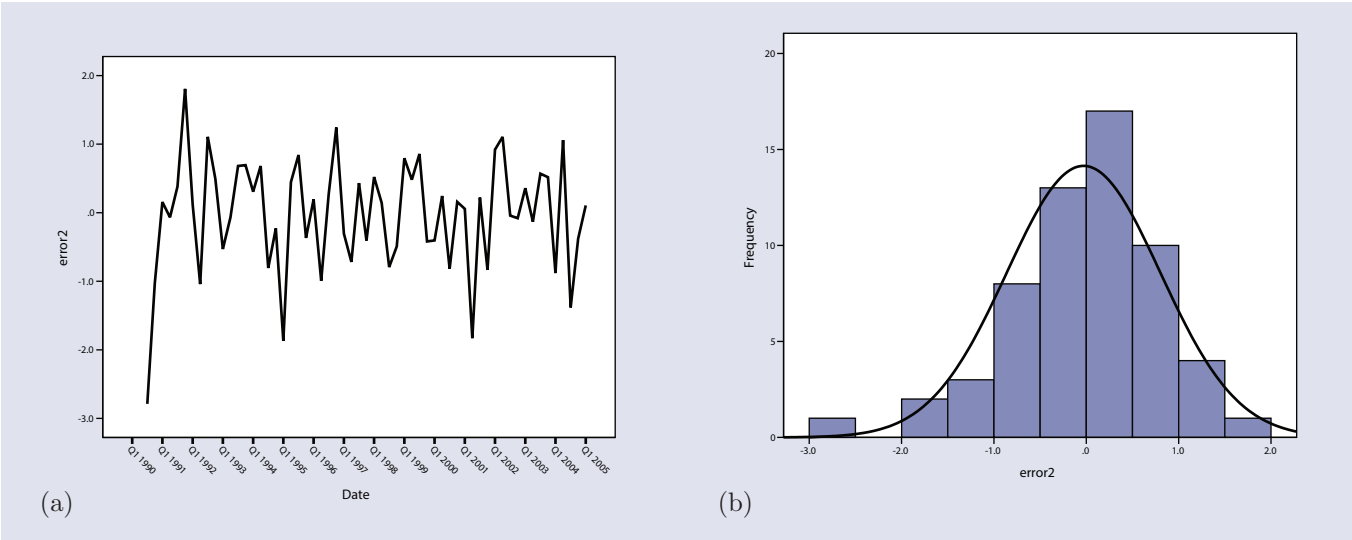


Figure S.28 Forecast errors: (a) time plot (b) histogram

The time plot shows that the forecast errors appear to be distributed around zero with constant variance. The histogram in Figure S.28(b) is less skewed than that in Figure 8.5(b), so it is more plausible that the distribution of the forecast errors is normal for the $ARIMA(1, 2, 0)$ model than for the $ARIMA(0, 2, 1)$ model.

The correlogram for the forecast errors is shown in Figure S.29.

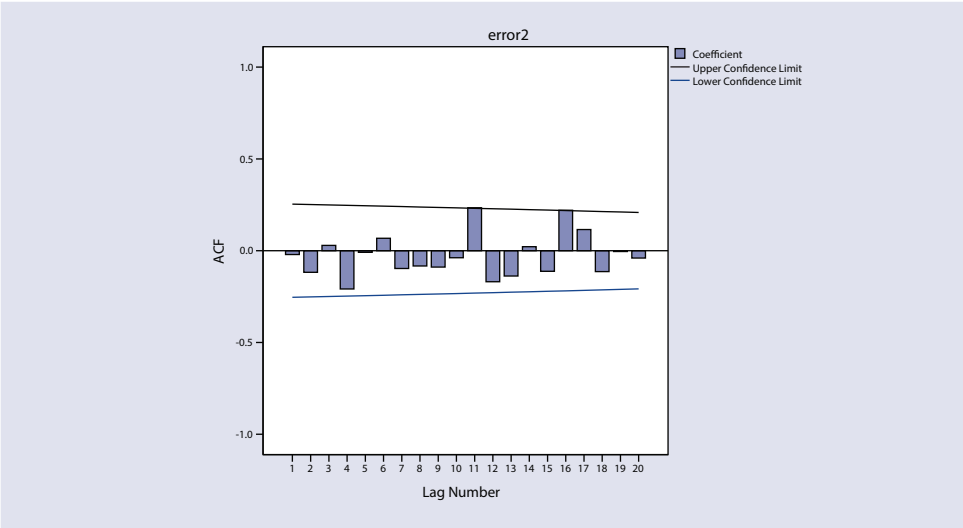


Figure S.29 Correlogram for forecast errors

This shows no evidence of non-zero autocorrelation at lags 1 to 20. The value of the Ljung–Box test statistic is 21.278, and $p = 0.381$. Hence there is little evidence against the null hypothesis of zero autocorrelation at lags 1 to 20.

In conclusion, the $ARIMA(1, 2, 0)$ model is adequate for this time series.

- (d) The $RMSE$ for the $ARIMA(0, 2, 1)$ model fitted in Computer Activity 8.4 was 0.826. Since an $RMSE$ of 0.840 was obtained with the $ARIMA(1, 2, 0)$ model, the $ARIMA(0, 2, 1)$ model fits slightly better. (The corresponding values of the SSE were 38.9 and 40.2.) On the other hand, the distribution of the forecast errors for the $ARIMA(1, 2, 0)$ model appears closer to a normal distribution. On balance, perhaps the $ARIMA(0, 2, 1)$ model is better, though either model is acceptable.

Solution 8.8

- (a) Applying a log transformation and fitting an ARIMA model is described in Computer Activity 8.7. The *RMSE* is given in the **Model Statistics** output table: $RMSE = 3\,227\,840.036$. Estimates of the parameters θ_1 and μ are given in the **ARIMA Model Parameters** table: $\hat{\theta}_1 = 0.869$, $\hat{\mu} = -0.002$.
- (b) The four plots required to check the fit and adequacy of the fitted model are shown in Figures S.30 and S.31. (The variables were ‘tidied up’ in the **Variable View** panel in the usual way before the plots were produced.)

Figures S.30(a) and (b) were produced with the **Sequence Charts** dialogue box

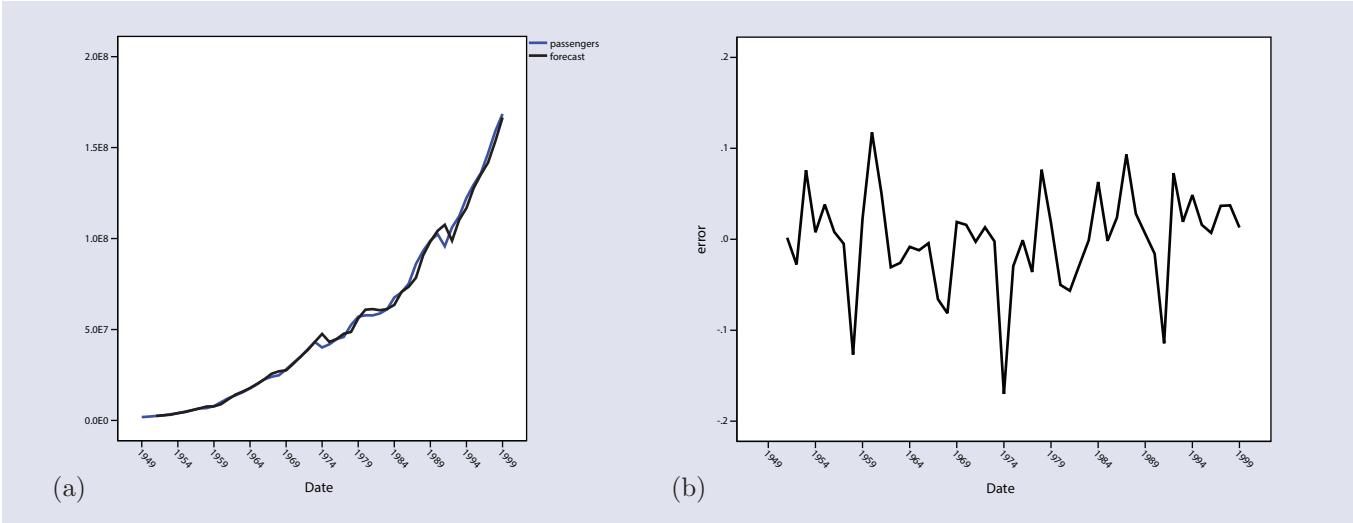


Figure S.30 (a) Observed and forecasted values (b) Time plot of forecast errors

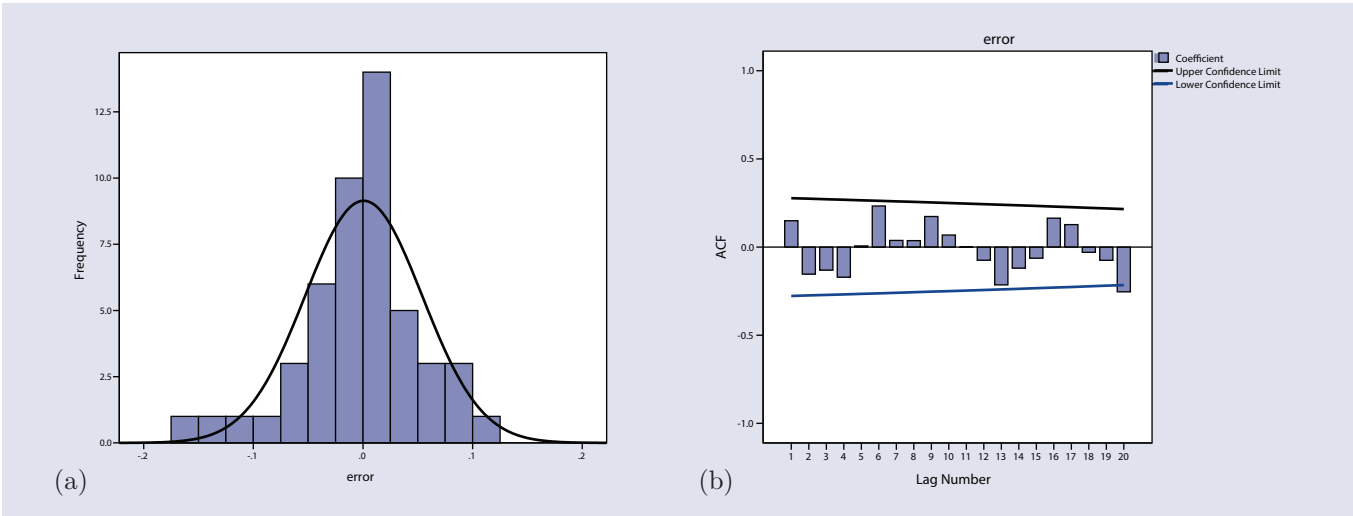


Figure S.31 (a) Histogram of forecast errors (b) Correlogram for forecast errors

The value of the Ljung–Box test statistic is 24.734, with $p = 0.212$.

- (c) The 1-step ahead forecasts provide a reasonable fit to the observed values.
- The time plot of forecast errors suggests that they are distributed with mean zero and constant variance. The histogram shows that the normal model is not unreasonable.

The correlogram shows an alternating pattern. However, all but one of the sample autocorrelations at lags 1 to 20 lie within the 95% significance bounds. (The exception is the autocorrelation at lag 20.) The Ljung–Box test provides little evidence against the null hypothesis of zero autocorrelation at lags 1 to 20.

In conclusion, the ARIMA(0,2,1) model is adequate for the time series of logarithms of annual numbers of airline passengers.

Solution 8.10

- (a) You fitted the model in Computer Activity 8.8 using the method described in Computer Activity 8.7. The forecasts and prediction limits are obtained using the method described in Computer Activity 8.9. In the **Options** panel of the **Time Series Modeler** dialogue box, select **First case after end of estimation period through a specified date** and enter 2010 in the **Year** field. The multiple plot required is shown in Figure S.32.

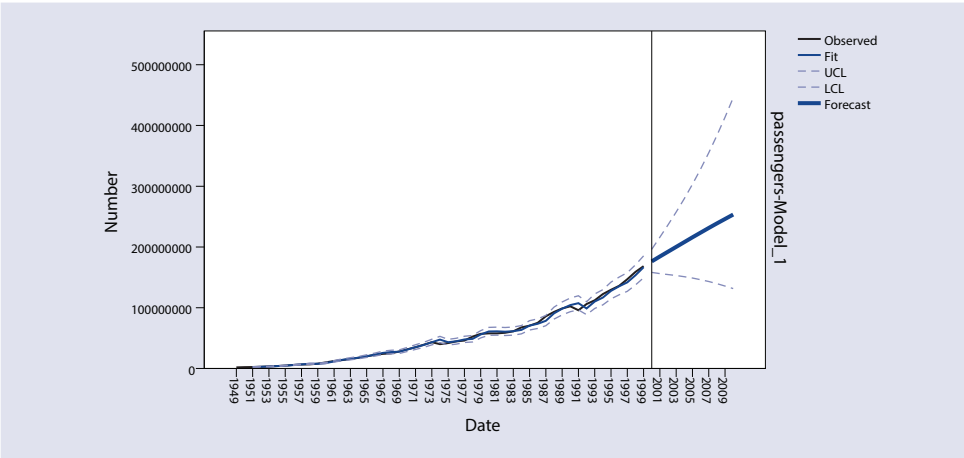


Figure S.32 Passenger numbers, forecasted values and prediction limits

- (b) The annual number of airline passengers in the UK is forecasted to increase from the observed value of 168.5 million in 1999 to 253.5 million in 2010, with 95% prediction interval (131.5, 445.7) million. The uncertainty surrounding the forecasts increases dramatically as predictions are made further into the future.
- (c) Since 1999 (the last year on which the forecasts are based), several factors may have affected air travel, for example: concerns over security following the attacks in the United States on 11 September 2001; increased oil prices following conflict in the Middle East and natural disasters; and pressure to reduce travel owing to concern over climate change.

Solutions to Computer Exercises

Solution 1

This exercise covers some of the ideas and techniques discussed in Sections 2, 4, 6 and 9 of *Book 2* and in Chapters 1, 2, 4 and 6 of this computer book.

(a) A time plot is shown in Figure S.33.

Obtaining a time plot and editing the labelling on the time axis are described in Computer Activity 1.4.

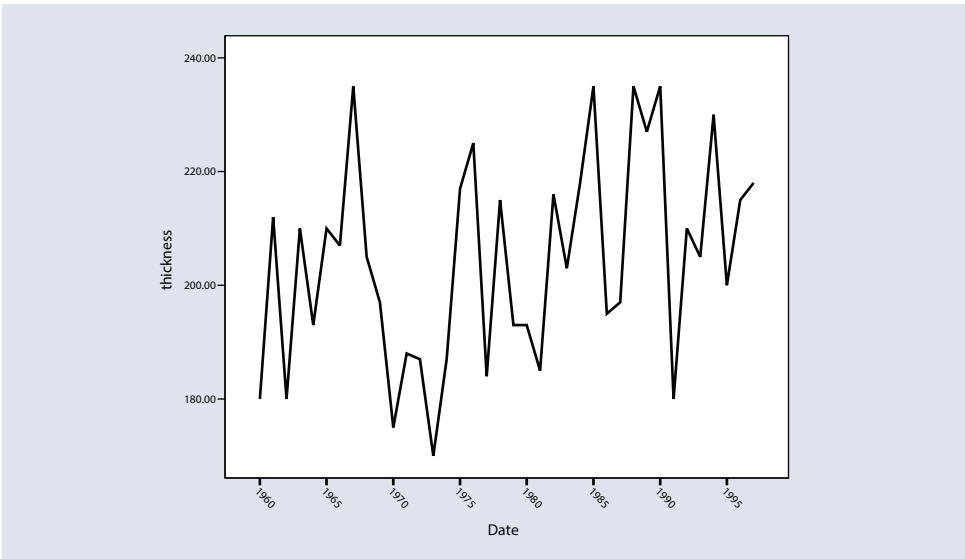


Figure S.33 Time plot of maximum spring ice thickness

The data are annual, so the time series has no seasonal component. There may be an increasing trend, though this is hard to see owing to the very large irregular fluctuations. There is little evidence that the size of these fluctuations varies with the level, so an additive model is appropriate.

(b) A multiple time plot of the data and the moving average is shown in Figure S.34.

The simple centred moving average and multiple time plot are obtained using the method described in Computer Activity 2.1.

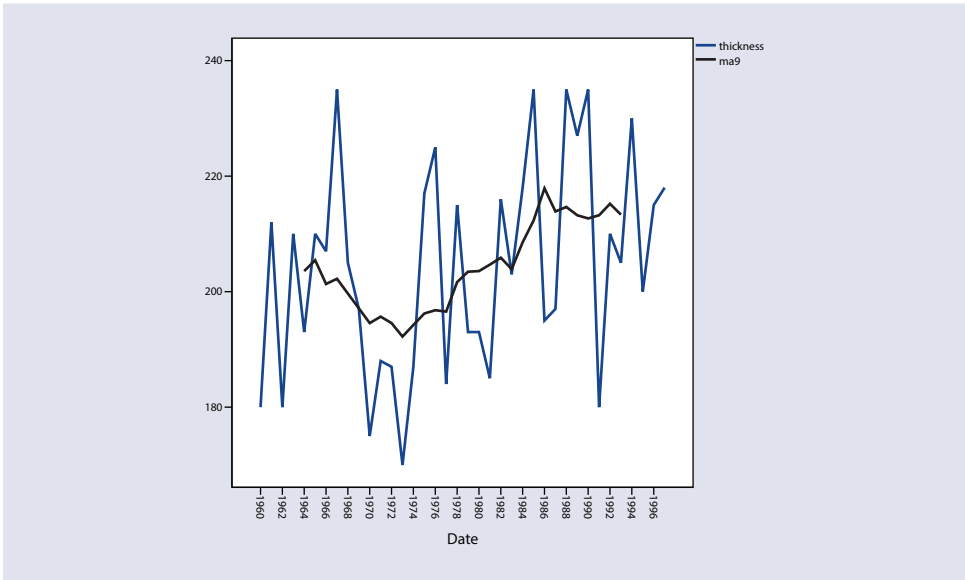


Figure S.34 Ice thickness and moving average of order 9

The underlying trend revealed by the moving average is as follows: the average ice thickness declined until about 1973, and increased thereafter.

- (c) The optimal value of the smoothing parameter α is 0.099. This is a small value, indicating that little weight is placed on the most recent observations. The $RMSE$ is 18.398. There are $n = 38$ observations and $k = 1$ smoothing parameters, hence

$$SSE = (38 - 1) \times 18.398^2 = 12\,523.9969 \dots \simeq 12\,524.$$

The predicted ice thickness for the Spring of 1998 is 210.37 cm (to two decimal places).

- (d) The plots required are shown in Figures S.35 and S.36.

Simple exponential smoothing in SPSS is described in Computer Activity 4.3.

Obtaining a correlogram is described in Computer Activity 6.1, and checking normality in Computer Activity 6.3.

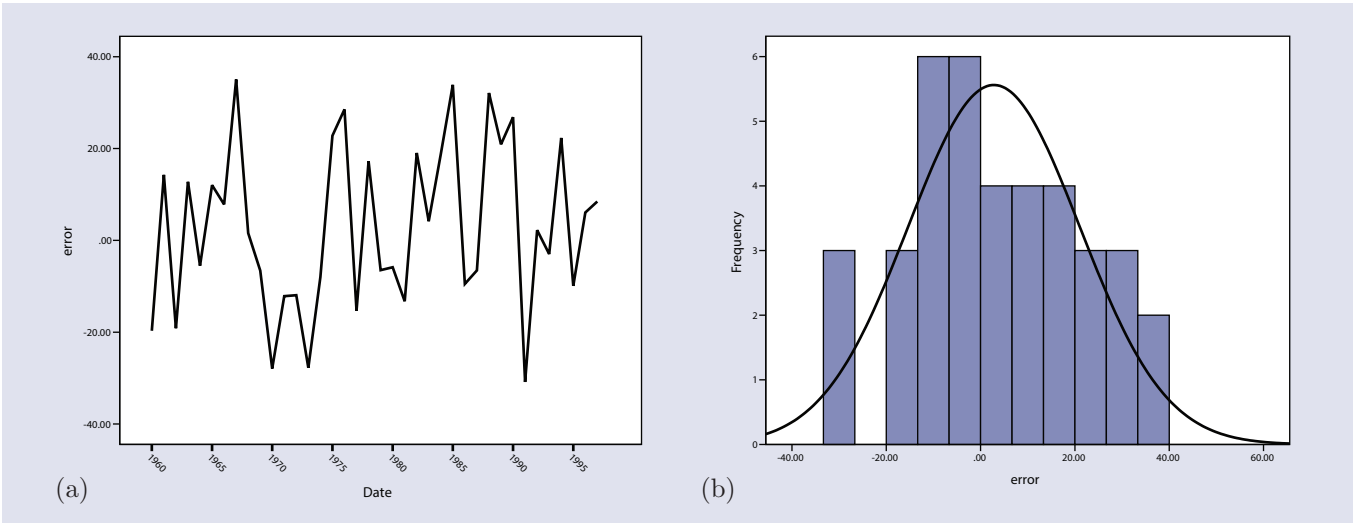


Figure S.35 Forecast errors: (a) time plot (b) histogram

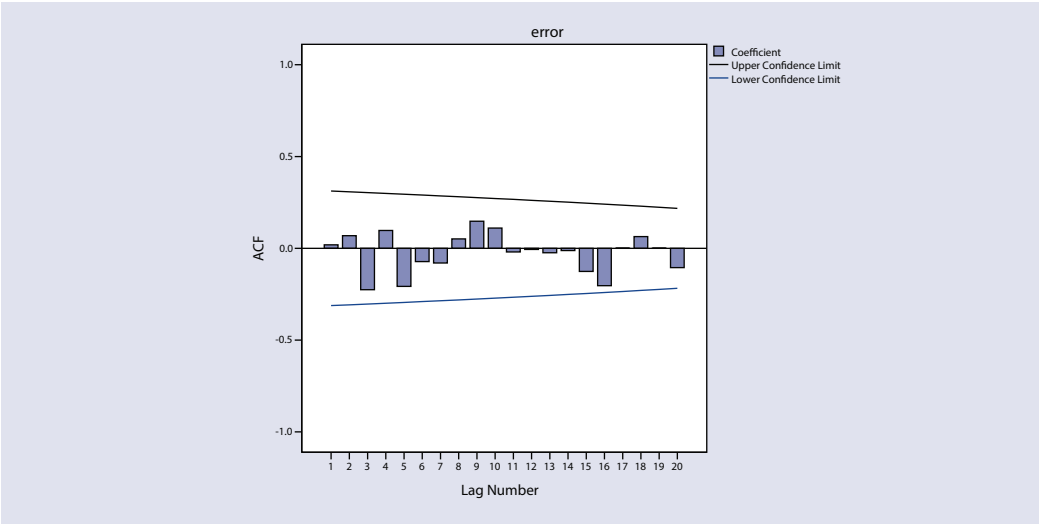


Figure S.36 Correlogram for forecast errors

The time plot does not reveal any systematic variation in level or variance of the forecast errors, which appear to be distributed about zero. The histogram and superimposed normal curve are not all that close. However, the differences are not great for a data set as small as this one, so the errors may well be normally distributed. The correlogram provides little evidence of non-zero autocorrelation at lags 1 to 20. The value of the Ljung–Box test statistic is 12.543, with $p = 0.896$. Thus there is little evidence against the null hypothesis of zero autocorrelation at lags 1 to 20. In conclusion, it is reasonable to assume that the forecast errors are white noise.

- (e) Since the SSE is 12 524 and there are 38 observations, the 95% prediction limits are as follows:

$$210.37 - 1.96 \times \sqrt{12\,524/38} \simeq 174.79,$$
$$210.37 + 1.96 \times \sqrt{12\,524/38} \simeq 245.95.$$

The maximum thickness of spring ice in Resolute Bay has tended to increase between 1973 and 1997, with much year-to-year variation. The forecasted maximum ice thickness for the spring of 1998 is about 210 cm, with 95% prediction interval (175, 246).

Solution 2

This exercise covers some of the ideas and techniques discussed in Sections 4 and 7 of *Book 2* and in Chapters 1, 2, 3 and 5 of this computer book.

- (a) The time plot with edited labelling on the time axis is shown in Figure S.37.

See Computer Activity 6.4.

The 95% prediction interval calculated by SPSS is (173, 248).

Obtaining a time plot and editing the time axis labels are described in Computer Activity 1.4.

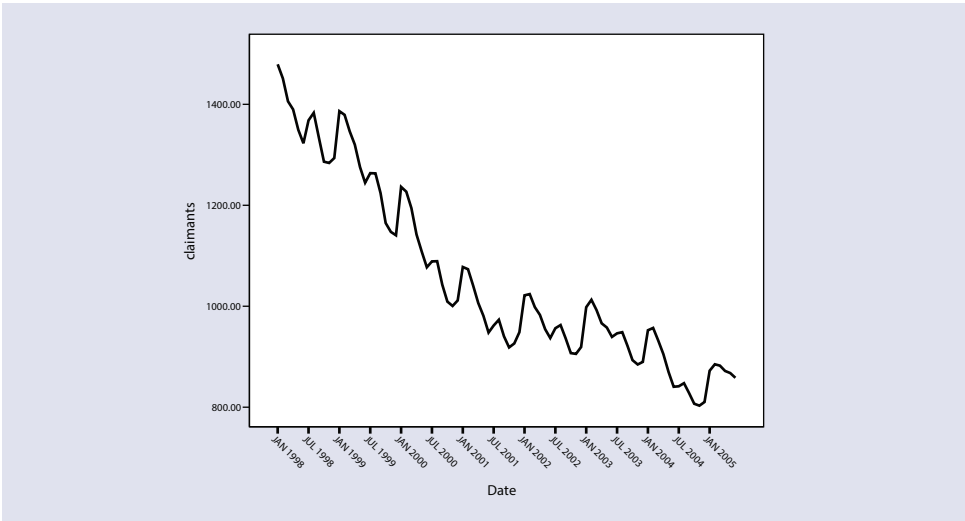


Figure S.37 Time plot of monthly claimant count

It is apparent (using the labels on the time axis) that the time series is seasonal with period one year. The size of the seasonal fluctuations does not vary much with the level of the series. There is a declining trend. The irregular component is not apparent, owing to the scale on which the plot is drawn. Thus as far as it is possible to see, there is no evidence against an additive model being appropriate for this time series.

- (b) Estimating the seasonal factors of a time series is discussed in Computer Activity 3.1. The estimated seasonal factors are given in the **Seasonal Factors** output table, which is reproduced in the margin.

The largest positive seasonal factor is the one for February; this is when the number of claimants reaches its seasonal high. The largest negative seasonal factor is the one for October, when the claimant count reaches its seasonal minimum. The seasonal factors for October to December are negative, perhaps because this period corresponds to the run up to Christmas each year, when more seasonal work is available.

Seasonal Factors	
Series Name: claimants	
Period	Seasonal Factor
1	53.90243
2	61.41354
3	40.51701
4	17.04826
5	-5.59826
6	-25.60174
7	-4.35313
8	8.47426
9	-19.32455
10	-47.62396
11	-46.67217
12	-32.18170

- (c) Obtaining a moving average is described in Computer Activity 2.1. Using this to obtain an estimate of the irregular component is described in Computer Activity 3.3. The time plot of the estimated irregular component is shown in Figure S.38.

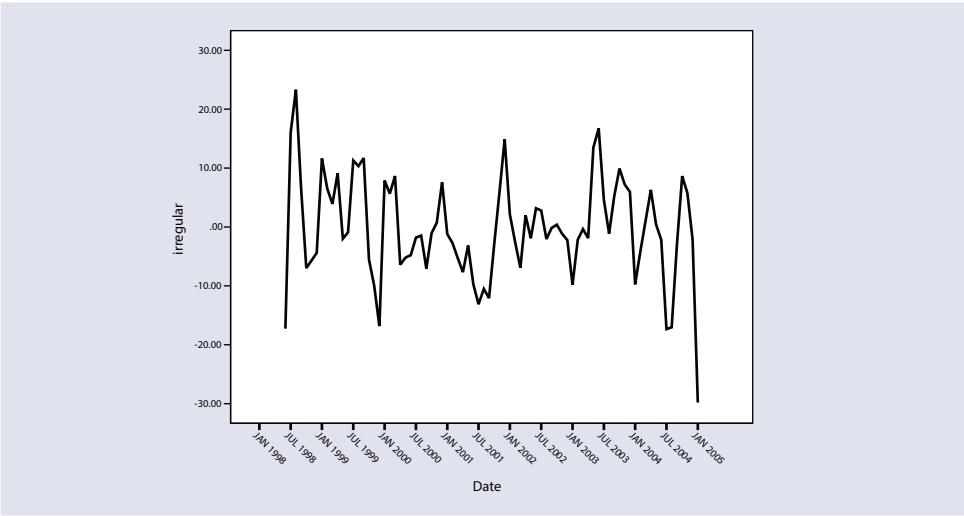


Figure S.38 Estimated irregular component

- (d) The optimal parameter values are $\alpha = 0.600$, $\gamma = 0.201$ and $\delta = 1.000$. Thus the most recent observations have a moderate influence on the level, little influence on the trend, and a great deal of influence on the seasonal component.
- (e) A time plot of observed and forecasted values is shown in Figure S.39.

Applying the Holt–Winters method in SPSS is described in Computer Activity 5.3.

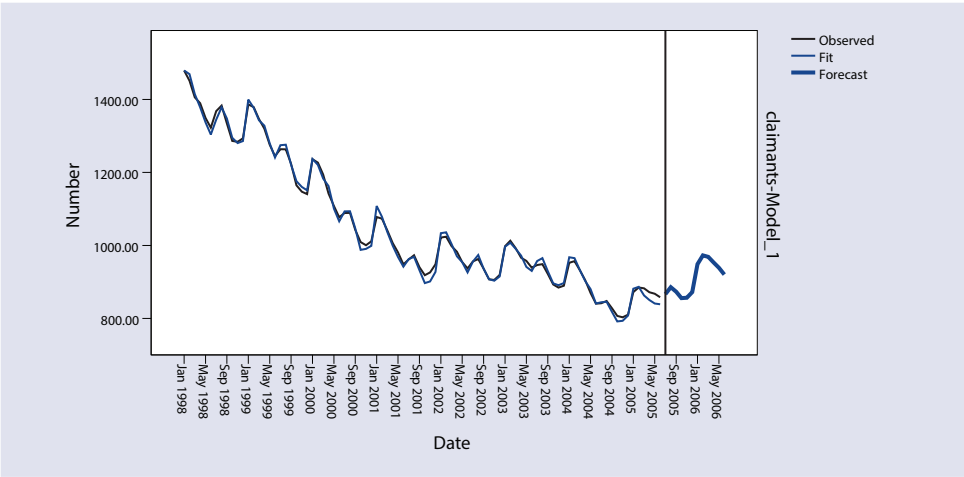


Figure S.39 Observed and forecasted numbers of claimants

The claimant count is predicted to rise between June 2005 and June 2006, with the same seasonal pattern as observed in previous years.

Solution 3

This exercise covers some of the ideas and techniques discussed in Part III of *Book 2* and in Chapters 7 and 8 of this computer book.

- (a) A time plot of the seasonally adjusted time series `saflow` is shown in Figure S.40.

Obtaining a time plot is described in Computer Activity 1.4.

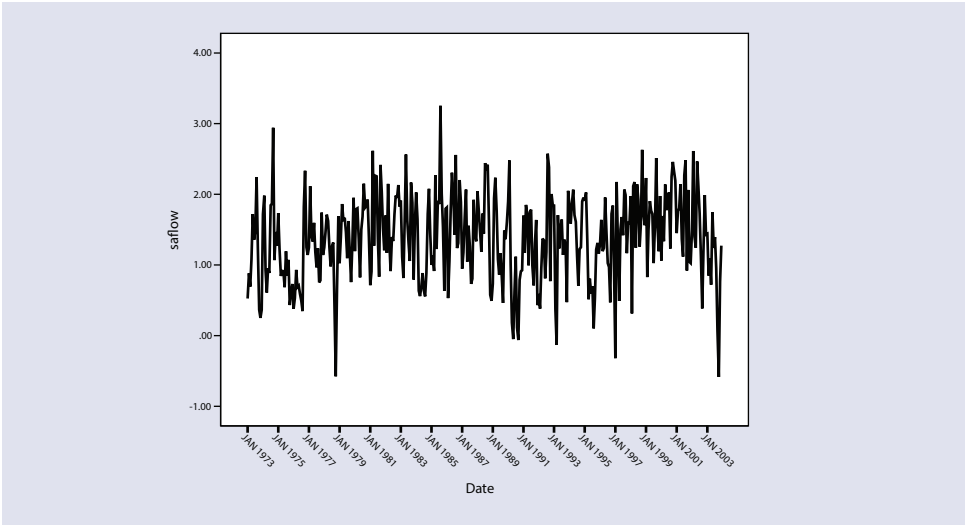


Figure S.40 Time plot of `saflow`

There is no systematic variation in level or in the size of the fluctuations, so the series is stationary in mean and in variance. There is no reason to believe it is not stationary in correlation. Thus it is reasonable to conclude that the time series is stationary.

Stationarity is discussed in Chapter 7 and in Section 11 of *Book 2*.

- (b) The correlogram and the partial correlogram for the seasonally adjusted time series are shown in Figure S.41.

Obtaining correlograms and partial correlograms is described in Computer Activity 8.1.

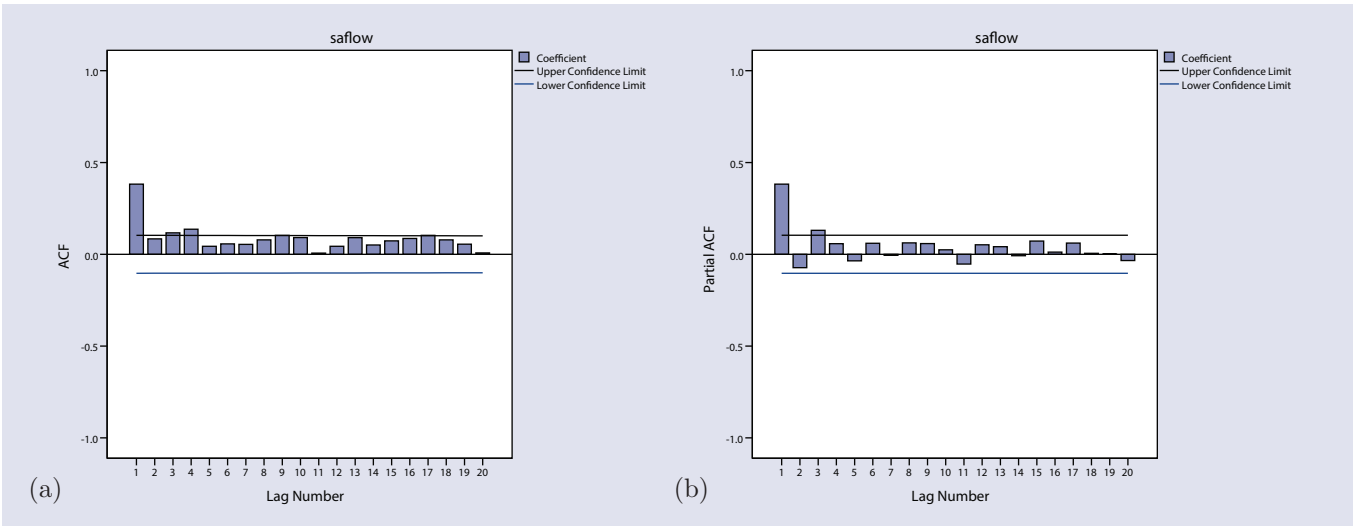


Figure S.41 (a) Correlogram (b) partial correlogram

Both the correlogram and the partial correlogram show relatively large values at lag 1. Subsequent sample autocorrelations and partial autocorrelations are much smaller, and seldom exceed the significance bounds.

- (c) The large sample autocorrelation at lag 1 suggests an MA(1) model. The large sample partial autocorrelation at lag 1 suggests an AR(1) model. An alternative interpretation is that both the sample ACF and sample PACF decline exponentially, suggesting an ARMA(1, 1) model. Using the principle of parsimony, the ARMA(1, 1) model is set aside. This leaves the MA(1) and AR(1) models as plausible candidates. Since no differencing is required to produce stationarity, the two possible ARIMA models are ARIMA(1, 0, 0) and ARIMA(0, 0, 1).
- (d) The *RMSE* for the ARIMA(1, 0, 0) model is 0.564. The *RMSE* for the ARIMA(0, 0, 1) model is 0.563. There is little difference between these values, but the MA(1) model gives a very slightly better fit.

Fitting an ARIMA model and interpreting the parameter estimates are described in Computer Activity 8.4.

The model equations are as follows.

ARIMA(1, 0, 0):

$$X_t - 1.373 = 0.383(X_{t-1} - 1.373) + Z_t.$$

ARIMA(0, 0, 1):

$$X_t - 1.374 = Z_t + 0.401Z_{t-1}.$$

- (e) Checking the adequacy of a model involves checking that the forecast errors may be assumed to be white noise. The plots required to check the adequacy of the ARIMA(0, 0, 1) model are shown in Figures S.42 and S.43.

Checking the adequacy of a model is described in Computer Activity 8.5.

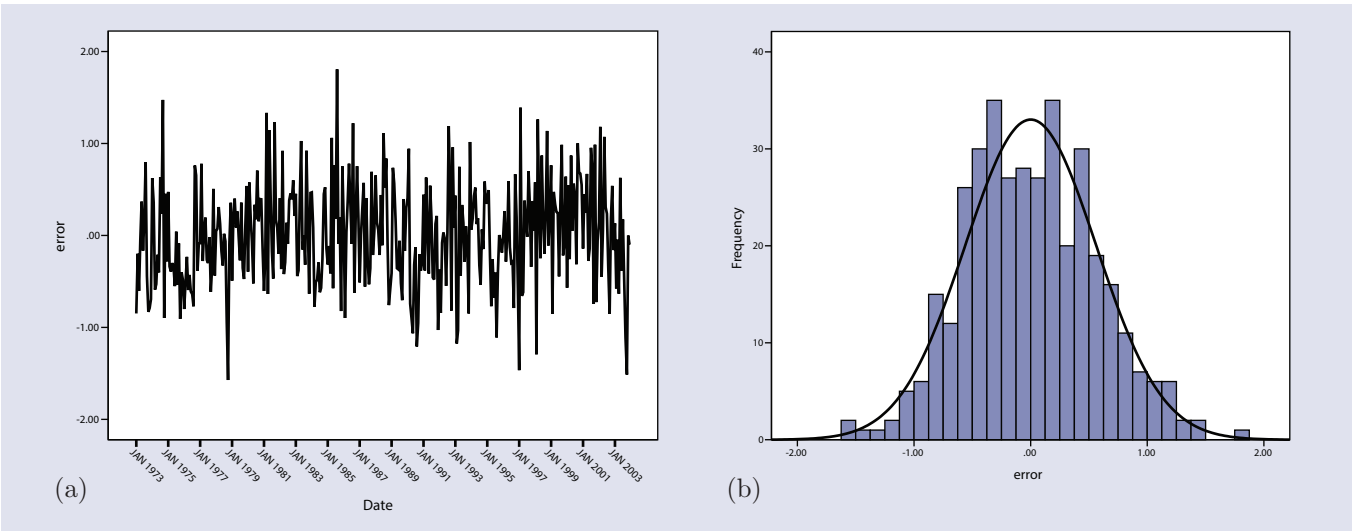


Figure S.42 Forecast errors: (a) time plot (b) histogram

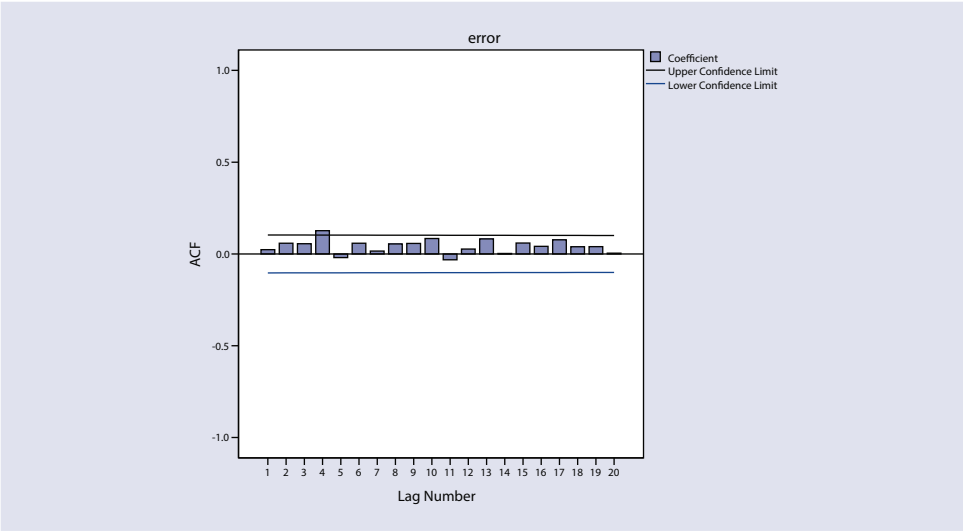


Figure S.43 Correlogram for forecast errors

From the time plot, the errors appear to fluctuate about zero with constant variance. The normal curve fits the histogram quite well, so the forecast errors seem to be normally distributed. The correlogram provides little evidence of non-zero autocorrelation. The value of the Ljung–Box test statistic is 24.314, and $p = 0.229$. Thus it is reasonable to assume that the forecast errors are white noise.

- (f) The forecasted value for December 2004 is 1.37, with 95% prediction interval (0.18, 2.57). The time plot is shown in Figure S.44.

Obtaining forecasts with ARIMA models is described in Computer Activity 8.9.

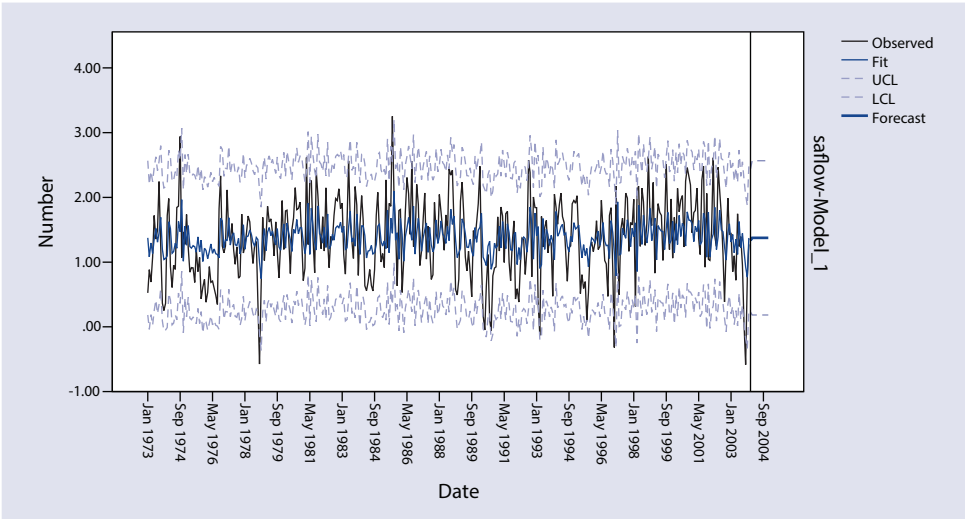


Figure S.44 Data, forecasted values and prediction intervals

The prediction intervals, both for the past data (up to December 2003) and the forecasts up to December 2004, are very wide.

Index

- ACF plots 34, 42
- ARIMA 45
- ARIMA modelling 42
- Autocorrelations** 33, 42
- Box–Ljung test 34
- Chart Editor** 7
- checking the model 47
- checking the normality assumption 36
- Compute Variable** 10
- correlogram 33, 42
- Create Models** 21, 45
- Create Time Series** 12
- Define Dates** 4
- differencing 38
- entering time series data 4
- estimate of the irregular component 19
- estimate of the trend component 19
- estimated seasonal factors 18
- Exponential Smoothing** 22
- fitting an ARIMA model 45
- Forecasting** 6
- forecasting 33
 - with ARIMA models 53
- Histogram** 36
- histograms and normal curves 36
- Holt’s exponential smoothing 27
- Holt–Winters exponential smoothing 30
- Label orientation** 7
- Labels & Ticks** 7
- Ljung–Box test 34
- log transformation 9
- maximum value 14
- minimum value 14
- moving averages 12
- multiple time plot 13
- 1-step ahead forecasts 23
- PACF plots 42
- Parameter estimates** 22
- Partial Autocorrelations** 42
- partial correlogram 42
- prediction intervals 53
- prediction limits 53
- Properties** 7
- RMSE* 22
- sample ACF 42
- sample PACF 42
- Scale** 14
- Seasonal Decomposition** 16
- seasonal factors 18
- seasonally adjusted series 18
- selecting an ARIMA model 42
- Sequence Charts** 6, 9
- significance bounds 34, 43
- simple centred moving average 12
- simple exponential smoothing 21
- smoothing parameter 21
- Span** 12
- square root transformation 10
- SSE* 21
- stationarity in mean 38
- stationarity in variance 38
- stationary time series 38
- time plot 6
- Time Series Modeler** 22
- transforming time series 8, 10
- trend-cycle component 19

